THIS WEEK

EDITORIALS

WORLD VIEW A political resolution for US scientists in 2013 **p.7**

NEUROSCIENCE Rita Levi-Montalcini dies at 103 **p.8**

POLIO Pakistan vaccinations on hold after string of shootings **p.8**

In search of credit

Explicit recognition of researchers' contributions to science is becoming more comprehensive. Not before time — especially as a means of crediting referees.

Last year, this journal received an unusual request: could three authors have it indicated in a footnote that they were joint second authors on a paper? We refused — for better or worse, our policy is to allow no more than three authors in first and last positions on a paper. But authorship order is a much greater obsession in some disciplines than others (the example in question was from biology). And there could hardly be a more clumsy way of indicating credit — not to mention the disputes that it provokes among co-authors.

For several years, *Nature* and the *Nature* research journals have insisted that each author's contribution should be indicated in a statement at the end of any paper. However, these statements are not systematic, and are not accompanied by metadata to make them more searchable. So although this approach works reasonably well in indicating who did what on a particular paper, there is potential for such statements to cumulatively provide a database of the skills and experience of individual researchers. Through such statements, it could become transparently clear that, say, John Smith was responsible for the development of a particular technique and had applied it in multiple contexts. At *Nature*, we are working on ways to increase the utility of author contribution statements and so achieve such transparency.

Of course, it would help to know which John Smith we are talking about. And here is where last year's launch of the Open Researcher and Contributor ID (ORCID) facility is to be welcomed. The core function of ORCID — a community collaboration (see go.nature.com/sy3qnp) — is to assign every researcher a number and a web page, thereby providing a unique identifier and so disambiguation. The web page enables the researcher to record their contributions: papers they have published and — a facility to come — their research grants and patents. *Nature* journals authors can link their ORCID to their account in our manuscript submission and tracking system, and we will soon be publishing authors' ORCIDs in papers. (Readers can register for ORCID here: https://orcid.org/register; see also *Nature* **485,** 564; 2012.)

In contrast to such public activities, refereeing tends to be a private affair, whether for funding agencies or for journals. But it is of immense value and deserves its own credit. Referees can examine a submission only for its surface validity rather than for its deeper truth, but that in itself involves a substantial commitment. Some may devote days to the task if they are sufficiently stimulated or worried. The more that can be done to reward such dedication the better.

That is why *Nature* and the *Nature* journals have introduced two ways in which referees can be given credit. Any referee who, in a given year, has refereed three or more papers for any of the journals will receive a letter acknowledging their contribution and a free subscription to their choice of one of the journals. More importantly, we have recently introduced a system by which our referees can download a statement of the number of papers they have refereed for us. This report is available by logging into the 'My Account' page on any *Nature* journal's manuscript submission and tracking system and reflects the refereeing activity across all *Nature* journals. If nothing else, such statements provide a formal reference that someone can pass on to employers, government agencies and others enlightened enough to appreciate the value of such contributions.

All of these developments are ways in which researchers can gain explicit credit for contributions that have previously relied more on word of mouth. This is a trend that we will continue to support and encourage. ■

Safety catch

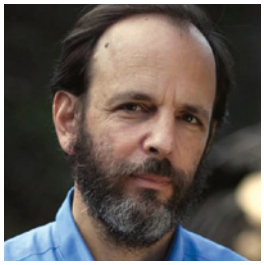
International laboratory survey offers comfort — and caution.

In Lake Wobegon, the fictional town invented by the US humorist Garrison Keillor, "all the women are strong, all the men are good-looking, and all the children are above average". In keeping with Keillor's gentle dig at the inflations of self-bias, if Lake Wobegon had research laboratories you can be sure that all the experiments would work, all the results would be significant and all the scientists would work safely.

This week, *Nature* reports the initial analysis of results from the first international survey of scientists' attitudes and behaviour towards lab safety, conducted by the University of California, Los Angeles, together with Nature Publishing Group (see page 9). The analysis hints at a Lake Wobegon bias in perceptions about safety: one-third of scientists say that safety is more important to them than it is to their colleagues, with only 2% voting the other way. Although most respondents say that their labs are safe places to work, they simultaneously report behaviour, such as frequent lone working, that seems to belie that confidence.

The survey was done to improve understanding of lab safety culture. Health-and-safety officers have long complained of a lack of international data. It would be premature to draw immediate conclusions from the quantitative results — for example, almost half the respondents reported being injured in the lab — because few other comparable data have been collected. But the results do caution against complacency.

So, as you return to your laboratories in the New Year, look around the benches, observe your own working practices and those of your colleagues, and evaluate your relationships with supervisors and safety officers. Not everyone can be above average — but awareness of how perception clashes with reality can help lift standards for all. ■



Science must be seen to bridge the political divide

Scientists in the United States are often perceived as a Democratic interest group. For science's sake this has to change, argues **Daniel Sarewitz**.

To prevent science from continuing its worrying slide towards politicization, here's a New Year's resolution for scientists, especially in the United States: gain the confidence of people and politicians across the political spectrum by demonstrating that science is bipartisan.

That President Barack Obama chose to mention "technology, discovery and innovation" in his passionate victory speech in November shows just how strongly science has come, over the past decade or so, to be a part of the identity of one political party, the Democrats, in the United States. The highest-profile voices in the scientific community have avidly pursued this embrace. For the third presidential election in a row, dozens of Nobel prizewinners in physics, chemistry and medicine signed a letter endorsing the Democratic candidate.

The 2012 letter argued that Obama would ensure progress on the economy, health and the environment by continuing "America's proud legacy of discovery and invention", and that his Republican opponent, Mitt Romney, would "devastate a long tradition of support for public research and investment in science". The signatories wrote "as winners of the Nobel Prizes in Science", thus cleansing their endorsement of the taint of partisanship by invoking their authority as pre-eminent scientists.

But even Nobel prizewinners are citizens with political preferences. Of the 43 (out of 68) signatories on record as having made past political donations, only five had ever contributed to a Republican candidate, and none did so in the last election cycle. If the laureates are speaking on behalf of science, then science is revealing itself, like the unions, the civil service, environmentalists and tort lawyers, to be a Democratic interest, not a democratic one.

This is dangerous for science and for the nation. The claim that Republicans are anti-science is a staple of Democratic political rhetoric, but bipartisan support among politicians for national investment in science, especially basic research, is still strong. For more than 40 years, US government science spending has commanded a remarkably stable 10% of the annual expenditure for non-defence discretionary programmes. In good economic times, science budgets have gone up; in bad times, they have gone down. There have been more good times than bad, and science has prospered.

In the current period of dire fiscal stress, one way to undermine this stable funding and bipartisan support would be to convince Republicans, who control the House of Representatives, that science is a Democratic special interest.

This concern rests on clear precedent. Conservatives in the US government have long been hostile to social science, which they believe tilts towards liberal political agendas. Consequently, the social sciences have remained poorly funded and politically vulnerable, and every so often Republicans threaten to eliminate the entire National Science Foundation budget for social science.

As scientists seek to provide policy-relevant knowledge on complex, interdisciplinary problems ranging from fisheries depletion and carbon emissions to obesity and natural hazards, the boundary between the natural and the social sciences has blurred more than many scientists want to acknowledge. With Republicans generally sceptical of government's ability and authority to direct social and economic change, the enthusiasm with which leading scientists align themselves with the Democratic party can only reinforce conservative suspicions that for contentious issues such as climate change, natural-resource management and policies around reproduction, all science is social science.

The US scientific community must decide if it wants to be a Democratic interest group or if it wants to reassert its value as an independent national asset. If scientists want to claim that their recommendations are independent of their political beliefs, they ought to be able to show that those recommendations have the support of scientists with conflicting beliefs. Expert panels advising the government on politically divisive issues could strengthen their authority by demonstrating political diversity. The National Academies, as well as many government agencies, already try to balance representation from the academic, non-governmental and private sectors on many science advisory panels; it would be only a small step to be equally explicit about ideological or political diversity. Such information could be given voluntarily.

To connect scientific advice to bipartisanship would benefit political debate. Volatile issues, such as the regulation of environmental and public-health risks, often lead to accusations of 'junk science' from opposing sides. Politicians would find it more difficult to attack science endorsed by avowedly bipartisan groups of scientists, and more difficult to justify their policy preferences by scientific claims that were contradicted by bipartisan panels.

During the cold war, scientists from America and the Soviet Union developed lines of communication to improve the prospects for peace. Given the bitter ideological divisions in the United States today, scientists could reach across the political divide once again and set an example for all. ■

POLITICIANS WOULD FIND IT MORE DIFFICULT TO ATTACK SCIENCE ENDORSED BY BIPARTISAN GROUPS OF SCIENTISTS.

➲ NATURE.COM
Discuss this article online at:
go.nature.com/fiw8gs

Daniel Sarewitz *is co-director of the Consortium for Science, Policy and Outcomes at Arizona State University, and is based in Washington DC.*
e-mail: daniel.sarewitz@asu.edu

SEVEN DAYS *The news in brief*

POLICY

Polio setback

Pakistan has partly suspended a three-day vaccination campaign led by the Global Polio Eradication Initiative that focused on the country's polio hotspots, after gunmen killed nine local health workers between 17 and 19 December last year. Pakistan is one of only three countries where polio has not been eradicated. On 1 January this year, seven health workers and teachers working for a Pakistan charity were also killed by gunmen in northwest Pakistan. No organization has yet claimed responsibility for the shootings, but militants linked to the Taliban are suspected. See go.nature.com/xzsc4s for more.

Fisheries progress

Commercial fishing in Europe is closer to being set on a more scientific footing, after members of the European Parliament's fisheries committee voted through reforms to the European Union's fisheries policies. The changes include the stipulation that levels for catches be set on the basis of scientific advice. The proposals must next be voted on by the whole parliament and then be agreed on by ministers. See go.nature.com/hcm4yv for more.

TB drug approval

The US Food and Drug Administration has for the first time approved a medicine to treat multidrug-resistant tuberculosis (TB). Bedaquiline was approved on 28 December as part of a combination therapy. Made by Janssen Therapeutics of Titusville, New Jersey, the drug works by inhibiting a mycobacterial ATP synthase enzyme. For more on forthcoming TB therapies, see page 14 and *Nature* **487**, 413–414 (2012).

Science in Korea

Park Geun-hye, elected South Korea's first female president on 19 December, has promised to make science a cornerstone of the government's policy. Park's conservative Saenuri party has also said that it will increase spending on research and development from 4% (in 2011) to 5% of the nation's gross domestic product by 2017. The proportion of that money going to basic research will also rise from 35.2% to 40% over the same period, Park said. See go.nature.com/on2lqh for more.

PEOPLE

EPA chief quits

US Environmental Protection Agency (EPA) administrator Lisa Jackson will step down early this year after four years in the job, she announced on 27 December. Jackson championed tighter pollution standards and worked to rein in greenhouse-gas emissions. It is unclear who will take over her position, but deputy EPA administrator Robert Perciasepe is likely to step in until the next appointee is approved. See go.nature.com/jzaiyn for more.

Nobel laureate dies

Rita Levi-Montalcini, a Nobel-prizewinning neuroscientist who became a national heroine in her home country of Italy, died on 30 December, aged 103. Levi-Montalcini shared the 1986 Nobel Prize in Physiology or Medicine for her work on the discovery of nerve growth factor. In her last decades she was made

ALESSANDRA BENEDETTI/CORBIS

a senator for life in Italy's parliament, and created sparks by blocking legislation that might have been unfavourable to research. See go.nature.com/li2pjn for more.

Microbiologist dies

Carl Woese, a microbiologist who established that archaea are a third domain of life, died on 30 December, aged 84. In 1977, Woese and colleagues at the University of Illinois at Urbana-Champaign examined evolutionary relationships using ribosomal RNA sequences to show that prokaryotes comprise two different groups: bacteria and archaea. His theory was not accepted until the mid-1980s.

RESEARCH

Antarctic lake flop

UK scientists have abandoned their efforts to drill through 3 kilometres of ice to reach the subglacial Lake Ellsworth in Antarctica. The team from the British Antarctic Survey had planned to use high-pressure hot water to bore through the ice, but burned too much fuel trying to connect boreholes. The researchers say that they will try again in future Antarctic summers. See go.nature.com/xcclen for more.

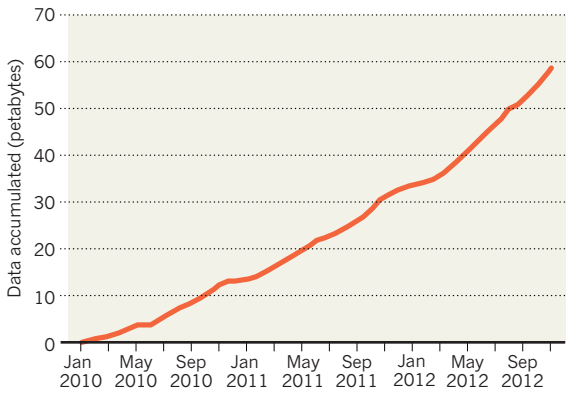
NATURE.COM
For daily news updates see:
www.nature.com/news

TREND WATCH

The Large Hadron Collider (LHC) ended 2012 with a bang, or rather 6 million billion billion bangs. That's the number of proton–proton collisions the accelerator has produced since it began running in 2010. Only 5 billion 'collisions of interest' were recorded, but that's still 60 petabytes of data — enough to fill the hard drives of around 80,000 laptop computers. The LHC will remain closed until 2015 for an energy upgrade. That will give physicists time to catch up with their analysis.

PARTICLE DATA PILE UP

Total amount of data accumulated by CERN's Large Hadron Collider since January 2010.



SOURCE: CERN

NEWS IN FOCUS

ARNO BURGI/DPA/PRESS ASSOCIATION

An international poll provides a lens into lab workers' attitudes to workplace welfare.

WORKPLACE

Safety survey reveals lab risks

Questionnaire suggests researchers not as safe as they feel.

BY RICHARD VAN NOORDEN

Scientists may have a false sense of security about the safety of their laboratories, according to early results from the first international survey of researchers' workplace attitudes and practices.

Some 86% of the roughly 2,400 scientists who responded said that they believe their labs are safe places to work. Yet just under half had experienced injuries ranging from animal bites to chemical inhalation, and large fractions noted frequent lone working, unreported injuries and insufficient safety training on specific hazards (see 'A question of safety').

"Understanding this disparity will be key to positively changing safety culture," says James Gibson, head of environmental health and safety at the University of California, Los Angeles (UCLA). The university's Center for Laboratory Safety, a research initiative set up in March 2011, commissioned the study as part of a wave of US-led efforts to examine safety culture following the shocking death of a 23-year-old research assistant, Sheharbano Sangji. She received horrific burns in a UCLA lab fire four years ago (see *Nature* http://doi.org/dnws3n; 2009), and her supervisor, organic chemist Patrick Harran, may face a criminal trial over her death. Other incidents, including a second lab death, at Yale University in New Haven, Connecticut, in 2011 (see *Nature* **472,** 270–271; 2011), have added to the concerns.

The study "is the most comprehensive attempt at gathering data on attitudes to safety that I've seen — and one more piece of information in a growing body of reports that point to the need to improve the culture around safety in our academic laboratories," says Dorothy Zolandz, director of the US National Academies Board on Chemical Sciences and Technology. Nature Publishing Group,the publisher of *Nature*, helped to launch the survey, as did the firm BioRAFT, which provides software for safety compliance and receives investment from Digital Science, a sister company to Nature Publishing Group. UCLA's Center for Laboratory Safety plans to analyse the data more closely later this year, but shared early results with *Nature*.

PART AND PARCEL

Some of the anonymized survey participants — who were mostly from the United States and United Kingdom, but also hailed from Europe, China and Japan — felt that any injuries they sustained were just part of the job. "Was scratched by a monkey," one scientist wrote. "It's bound to happen in that line of work, no matter how careful you are." Another was bitten while extracting venom from rattlesnakes; a third reported being sprayed on the face and hands with sulphuric acid, leading to US$3,000 of dermatology treatments. The most common injuries were minor — cuts, lacerations and needle pricks — but 30% of respondents said they had witnessed at least one 'major' lab injury, something that required attention from a medical professional. More than one-quarter of junior researchers said that they had experienced an injury that they hadn't reported to their supervisor.

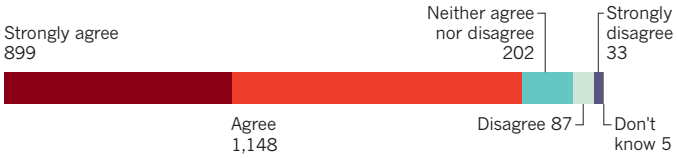
Yet the overwhelming majority of respondents asserted that their labs were safe places to work, that they had received sufficient safety training to minimize injury and that appropriate safety measures had been taken to protect employees. This level of comfort is similar to that found in other, smaller surveys, says Ralph Stuart, secretary of the American Chemical Society's health and safety division (which has conducted its own surveys on the matter).

But more specific questions in the survey reveal that safety standards are often not adhered to. Only 60% said they had received safety training on specific hazards or agents ▶

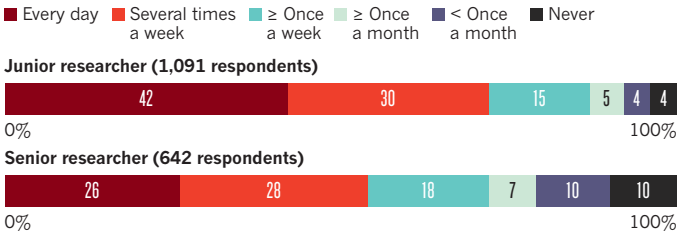
A QUESTION OF SAFETY

A survey of almost 2,400 scientists shows that although most believe their laboratories to be safe, about half have experienced injuries in the workplace. It also shows that junior and senior researchers have very different views of potentially hazardous practices.

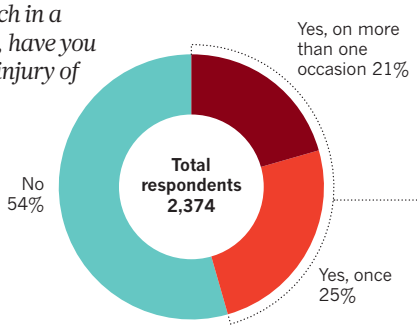
1 *To what extent do you agree or disagree with the following statement? "I feel that my lab is a safe place to work."*



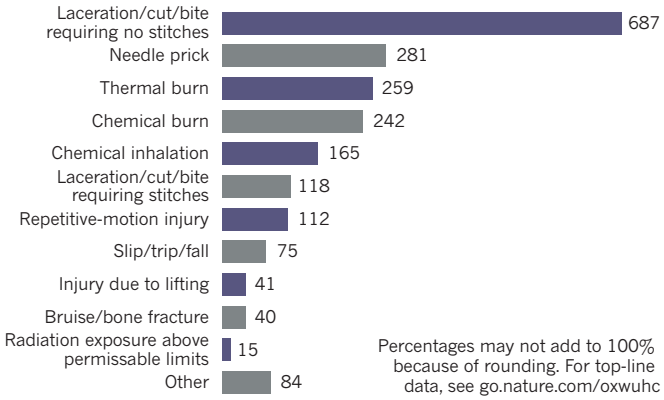
2 *In your lab, how frequently do people conduct experiments while working alone?*



3 *In the time that you've been conducting research in a laboratory setting, have you ever sustained an injury of any kind?*



4 *What was the nature of your injury or injuries?*



Percentages may not add to 100% because of rounding. For top-line data, see go.nature.com/oxwuhc

SOURCE: CENTER FOR LABORATORY SAFETY, UCLA/NPG/BONAMY FINCH

they worked with, and around half agreed that lab safety could be improved, with chemists (60%) most likely to feel this, and neuroscientists (30%) significantly less so.

OLD VERSUS YOUNG

One of the biggest gulfs picked up by the survey was differences in attitudes to safety between those in junior roles (such as postdocs and PhD students) and those in more senior positions (such as professors, heads of department and principal investigators). Around 40% of junior scientists said that people worked alone in their lab every day — compounding the risk to health should an accident occur — compared with just 26% of senior respondents (see graph 2), raising the possibility that supervisors are not always aware of the safety culture in their own group.

Overall, about two-thirds of researchers said that people worked alone in their lab at least several times a week. And only 12% of younger scientists said that safety was "paramount, and takes precedence over all other lab priorities", compared with 36% of senior scientists.

Younger researchers may have a clearer view of safety practices: controlling for other factors, junior researchers worked longer hours at the bench than their bosses. More than half of juniors worked over 40 hours per week, compared with just one-fifth of seniors, with almost 150 people overall reporting more than 60 hours per week.

Another finding — which comes as no surprise to health and safety experts — was the difference in how US and UK scientists assess risks before they start an experiment, which is, in part, a consequence of differences in legal requirements. Almost two-thirds of British scientists said that they used their organization's approved form for risk assessments — which is mandated by the nation's Health and Safety Executive — compared with only one-quarter of Americans. More than half of US scientists instead said they assessed risk "informally".

The biggest barriers to improving safety in the lab were "time and hassle" and "apathy", scientists said. "If I could have selected apathy three times over, I would have," one scientist wrote. These factors were closely followed by lack of understanding of safety requirements, lack of leadership and a focus on compliance requirements over safety. "Compliance does not equal to (sic) safety. More paperwork does not equal a safer lab; if anything, it makes it less safe," wrote one researcher. Another complained: "Safety training is very obviously aimed at instituting blind compliance to avoid liability. It is not aimed at teaching lab workers about why each safety measure is put in place."

Those feelings might explain researchers' mixed attitudes to the value of safety training, inspections and safety rules. Two-thirds of those surveyed thought that lab inspections improved safety, with senior scientists significantly more likely to agree than juniors. Yet two-fifths felt that safety training "focused on training compliance regulations rather than on improving laboratory safety", although 32% disagreed. And close to one-fifth of researchers said that lab safety rules had negatively impacted their lab productivity.

"These respondents are wrong, and this is a reflection of an urban myth [about the value of safety procedures] — it is highly frustrating," comments Neal Langerman, who runs the consulting company Advanced Chemical Safety, based in San Diego, California.

Some health and safety experts think that the survey — which involved almost 100 questions — was too broad and unfocused to draw definite conclusions. They also criticized its non-randomized sampling technique: the survey was sent out by e-mail to scientists who had registered on nature.com, and to research leaders, who were encouraged to pass it on to their lab scientists. But the experts acknowledged that it was a necessary and useful starting point for further investigation.

"This survey is a baseline study that leaves more questions than answers, but a perception survey is supposed to raise questions that need to be looked at," said Lou DiBerardinis, head of health and safety at the Massachusetts Institute of Technology (MIT) in Cambridge. DiBerardinis is on one of four teams to receive seed funding in 2012 from the Center for Laboratory Safety to study safety. He is working on a project led by MIT anthropologist Susan Silbey to track changing safety cultures by monitoring inspection records over seven years.

Zolandz says that this year, the National Academies Board on Chemical Sciences and Technology will team up with behavioural scientists to develop practical guidance for researchers on how to establish a better safety culture. In the various efforts that have followed Sangji's death, "that's one piece of the puzzle that's been missing", she says. "How do you get people to buy into safety?" ■ SEE EDITORIAL P.5

RESEARCH

New year, new science

Nature *looks ahead to the key findings and events that may emerge in 2013.*

STEM-CELL TRIALS

Landmark results from an early-stage clinical trial using human embryonic stem cells (hESCs) should appear this year. Biotechnology firm Advanced Cell Technology of Santa Monica, California, is injecting hESC-derived retinal cells into the eyes of around three dozen people with two forms of non-treatable degenerative blindness. It is the only company currently testing hESC therapies with US Food and Drug Administration (FDA) approval, and it hopes that the agency will give it the green light to test stem cells induced from adult cells in patients this year.

DIAGNOSTICS CONTROVERSY

The American Psychiatric Association will publish the fifth edition of its *Diagnostic and Statistical Manual of Mental Disorders* (*DSM-5*) in May, the first major update in 19 years to the standard reference guide for diagnosing mental illnesses. It will lead to controversial changes in clinical and research protocols, including restructured diagnoses for autism and major depression, although as a 'living document' the *DSM-5* will see further revisions.

CLIMATE ASSESSMENT

Climate scientists have spent years preparing the fifth assessment report from the Intergovernmental Panel on Climate Change, its first update since 2007. Part of that report is due to appear in September: the conclusions of Working Group I, which summarizes the basic science of global warming. In the United States, the Global Change Research Program's second assessment will detail the national impacts of climate change.

THE BIG BANG'S GLOW

One of the stunning images of the year could be provided by the comet ISON, which will pass close to the Sun in November and could outshine the full Moon as its surface boils away into space. Just as spectacular will be the Planck space telescope's map of the faint microwave afterglow from the Big Bang, which could even reveal ripples from gravitational waves generated during an initial period of cosmic 'inflation'. In other missions, NASA's LADEE spacecraft will orbit the Moon to study lunar dust; its MAVEN mission will launch to explore Mars' upper atmosphere; and the Curiosity rover will continue to send back results from the red planet's surface. Back on Earth, the massive 66-dish Atacama Large Millimeter/submillimeter Array in Chile will be completed.

➲ NATURE.COM
For a look back at the highlights of 2012, see:
www.nature.com/2012

DIET, MICROBES AND CANCER

Scientists increasingly suspect that our intestinal zoo of microbes might be the key link between diet and diseases such as cancer. A study last year connected a higher-than-normal proportion of the bacterium *Escherichia coli* to colorectal cancer in mice with inflammatory bowel disease (J. C. Arthur *et al. Science* **338,** 120–123; 2012). More studies this year will unpick the effect of diets on the gut microbiome and their implications for disease risk. Meanwhile, GlaxoSmithKline should find out whether the FDA approves its melanoma treatment trametinib, potentially the first in a new class of compounds that inhibit a kinase signalling pathway regulating cell growth.

MATT KAPUST/SANFORD LAB/AP

The Large Underground Xenon detector is seeking dark matter.

PARTICLE SEEKING

After contradictory sightings of dark-matter particles from various underground experiments, the Large Underground Xenon detector at the Sanford Underground Research Facility in Lead, South Dakota, may this year boost or rule out some of the claims. The king of particle hunters — the Large Hadron Collider at CERN near Geneva — will shut down until 2015 for an upgrade to enable more powerful collisions, but physicists will continue to pore over the data collected so far for hints of supersymmetry.

THE LOWER DEPTHS

Data will start flowing from the first completed segments of a giant underwater surveillance network, the US Ocean Observatories Initiative, which will cost US$386 million to build and will be completed by March 2015. It will monitor everything from undersea earthquakes and the effects of climate change on ocean circulation, to shifting ecosystems and ocean chemistry — all the way from the air to the seabed at seven sites around the globe. Meanwhile, British, American and Russian teams will be hoping to find out what kind of life, if any, exists in three deep, subglacial Antarctic lakes.

MAGICAL MATERIALS

Samarium hexaboride might be the next star of materials science, following hints last year that it is a topological insulator — conducting electricity on its surface, but behaving as an insulator inside. Graphene will remain a major celebrity, so expect a flood of reports about copycat materials such as boron nitride, tantalum disulphide and other two-dimensional sheets that can be stacked or sandwiched in precise layers.

GENES IN COURT

The US Supreme Court could decide a number of cases with science implications in 2013. It will re-examine whether genes are patentable as part of a three-year lawsuit considering the validity of patents held by Myriad Genetics in Salt Lake City, Utah. It may also rule on a challenge to seed firm Monsanto, based in St Louis, Missouri, from a farmer who wants to plant seeds gathered from previously grown genetically modified crops, rather than buying new stock from the company. And the court will consider whether brand-name pharmaceutical companies can pay generic makers to delay their launch of generic drugs.

PAPER MONEY

A UK policy that requires publicly funded researchers to make their results freely available will take effect from April. Other countries could soon follow — a Global Research Council meeting is set to discuss the matter in May. But many scientists will be worrying more about budgets, with the United States considering drastic spending cuts that could take effect early this year (see page 13), and Europe set to continue debating the proposed €80 billion (US$104 billion) in funding for its 2014–20 research programme, Horizon 2020. ■

COMPILED BY RICHARD VAN NOORDEN

ENERGY

Methane leaks erode green credentials of natural gas

Losses of up to 9% show need for broader data on US gas industry's environmental impact.

BY JEFF TOLLEFSON

Scientists are once again reporting alarmingly high methane emissions from an oil and gas field, underscoring questions about the environmental benefits of the boom in natural-gas production that is transforming the US energy system.

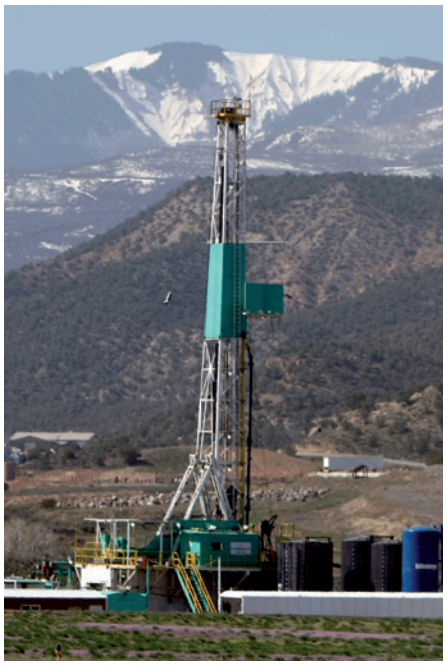
The researchers, who hold joint appointments with the National Oceanic and Atmospheric Administration (NOAA) and the University of Colorado in Boulder, first sparked concern in February 2012 with a study[1] suggesting that up to 4% of the methane produced at a field near Denver was escaping into the atmosphere. If methane — a potent greenhouse gas — is leaking from fields across the country at similar rates, it could be offsetting much of the climate benefit of the ongoing shift from coal- to gas-fired plants for electricity generation.

Industry officials and some scientists contested the claim, but at an American Geophysical Union (AGU) meeting in San Francisco, California, last month, the research team reported new Colorado data that support the earlier work, as well as preliminary results from a field study in the Uinta Basin of Utah suggesting even higher rates of methane leakage — an eye-popping 9% of the total production. That figure is nearly double the cumulative loss rates estimated from industry data — which are already higher in Utah than in Colorado.

"We were expecting to see high methane levels, but I don't think anybody really comprehended the true magnitude of what we would see," says Colm Sweeney, who led the aerial component of the study as head of the aircraft programme at NOAA's Earth System Research Laboratory in Boulder.

Whether the high leakage rates claimed in Colorado and Utah are typical across the US natural-gas industry remains unclear. The NOAA data represent a "small snapshot" of a much larger picture that the broader scientific community is now assembling, says Steven Hamburg, chief scientist at the Environmental Defense Fund (EDF) in Boston, Massachusetts.

The NOAA researchers collected their data in February as part of a broader analysis of air pollution in the Uinta Basin, using ground-based equipment and an aircraft to make detailed measurements of various pollutants, including methane concentrations. The researchers used atmospheric modelling to calculate the level of methane emissions required to reach those concentrations, and then compared that with industry data on gas production to obtain the percentage escaping into the atmosphere through venting and leaks.



Natural-gas wells such as this one in Colorado are increasingly important to the US energy supply.

J. DAVID ZALUBOWSKI/AP

The results build on those of the earlier Colorado study[1] in the Denver–Julesburg Basin, led by NOAA scientist Gabrielle Pétron (see *Nature* **482,** 139–140; 2012). That study relied on pollution measurements taken in 2008 on the ground and from a nearby tower, and estimated a leakage rate that was about twice as high as official figures suggested. But the team's methodology for calculating leakage — based on chemical analysis of the pollutants — remains in dispute. Michael Levi, an energy analyst at the Council on Foreign Relations in New York, published a peer-reviewed comment[2] questioning the findings and presenting an alternative interpretation of the data that would align overall leakage rates with previous estimates.

Pétron and her colleagues have a defence of the Colorado study in press[3], and at the AGU meeting she discussed a new study of the Denver–Julesburg Basin conducted with scientists at Picarro, a gas-analyser manufacturer based in Santa Clara, California. That study relies on carbon isotopes to differentiate between industrial emissions and methane from cows and feedlots, and the preliminary results line up with their earlier findings.

A great deal rides on getting the number right. A study[4] published in April by scientists at the EDF and Princeton University in New Jersey suggests that shifting to natural gas from coal-fired generators has immediate climatic benefits as long as the cumulative leakage rate from natural-gas production is below 3.2%; the benefits accumulate over time and are even larger if the gas plants replace older coal plants. By comparison, the authors note that the latest estimates from the US Environmental Protection Agency (EPA) suggest that 2.4% of total natural-gas production was lost to leakage in 2009.

To see if that number holds up, the NOAA scientists are also taking part in a comprehensive assessment of US natural-gas emissions, conducted by the University of Texas at Austin and the EDF, with various industry partners. The initiative will analyse emissions from the production, gathering, processing, long-distance transmission and local distribution of natural gas, and will gather data on the use of natural gas in the transportation sector. In addition to scouring through industry data, the scientists are collecting field measurements at facilities across the country. The researchers expect to submit the first of these studies for publication by February, and say that the others will be complete within a year.

In April, the EPA issued standards intended to reduce air pollution from hydraulic-fracturing operations — now standard within the oil and gas industry — and advocates say that more can be done, at the state and national levels, to reduce methane emissions. "There are clearly opportunities to reduce leakage," says Hamburg. ■

1. Pétron, G. *et al. J. Geophys. Res.* **117,** D04304 (2012).
2. Levi, M. A. *J. Geophys. Res.* **117,** D21203 (2012).
3. Pétron, G. *et al. J. Geophys. Res.* (in the press).
4. Alvarez, R. A., Pacala, S. W. Winebrake, J. J., Chameides, W. L. & Hamburg, S. P. *Proc. Natl Acad. Sci. USA* **109,** 6435–6440 (2012).

POLICY

US fiscal deal leaves science vulnerable

Congress delays mandatory cuts to agencies.

BY HEIDI LEDFORD

Law-makers in Washington DC greeted the new year with a frantic deal meant to avert a fiscal crisis. But the bill that passed the Senate and the House in pre-dawn votes on 1 and 2 January keeps researchers on tenterhooks for at least another two months by delaying mandatory spending cuts that could threaten science funding.

The last-ditch effort aimed to stave off the effects of the 'fiscal cliff': a painful series of tax hikes and budget cuts, scheduled to take effect in 2013, that is meant to reduce the US budget deficit but could push the country's weak economy back into recession. The cuts, known as the sequester, could shrink federal support for research and development (R&D) by US$57.5 billion over the next five years.

Rancorous last-minute negotiations yielded a tenuous agreement to raise taxes for the wealthy, but deferred decisions on the sequester — an across-the-board reduction of about 8% in nondefence discretionary spending, with at least 9% carved from defence — by two months. "The bill isn't ideal," says Eleanor Dehoney, vice-president of public policy at Research!America, a science-advocacy group based in Alexandria, Virginia. "But it does give advocates more time to convince policy-makers that cutting the US investment in R&D is counterproductive."

Federal agencies such as the National Institutes of Health (NIH) and the National Science Foundation have not announced what cuts they would make in the event of a sequester, but many institutions have drawn up contingency plans in case there is a sudden plunge in government funding. Nancy Andrews, dean of the Duke University School of Medicine in Durham, North Carolina, says that the medical school may need to cut back on graduate admissions, freeze salaries and reduce faculty recruitments if the NIH takes a severe hit.

NATURE.COM
Read about Barack Obama's second-term challenges:
go.nature.com/kaaorm

The two-month delay to the sequester is a complication for Andrews, who is wrestling with decisions on faculty retention and recruitment that must be made by mid-January. "If we need to spend more to help current faculty members maintain their research programmes through funding gaps, it will be harder to provide start-up funding for new faculty members," she says.

The delay means that law-makers will debate the sequester at the same time as they tackle the overall federal budget. The US Congress's inability to agree on a 2013 budget last year led it to adopt a continuing resolution that allows the government to keep functioning at roughly 2012 funding levels. That resolution expires on 27 March. By then, the country will also have reached a cap on the amount of money that it can borrow.

"The bill isn't ideal. But it does give advocates more time to convince policy-makers that cutting US investment in R&D is counter-productive."

This timing may increase the momentum behind big spending cuts, cautions Michael Lubell, director of public affairs for the American Physical Society in Washington DC. But it may also foster a more thorough discussion of which programmes should be cut and by how much, he adds. Science is unlikely to emerge unscarred, says Lubell, but it may end up in better shape than it would have under the original sequestration plan. "That doesn't mean it's going to be wonderful," he adds. "It means it will be less bad." ■

OBITUARY

Maxine Clarke (1954–2012)

Publishing Executive Editor of Nature.

For those of us who were lucky enough to have Maxine as a colleague, it is so difficult to believe that she is no longer with us. She died last month after a battle with cancer that lasted several years. It was a battle fought with great personal courage and resilience, as well as with immense professionalism.

Given the history of her illness, Maxine's death should not have come as much of a surprise. Yet, despite knowing about her challenges, I and her closest colleagues experienced shock and disbelief on hearing that she had passed away. That, in itself, reflects her steadfast refusal to be defined by her illness. She instead maintained her active commitment to *Nature* and her colleagues — who equally refused to conceive that anything could deprive them of her counsel.

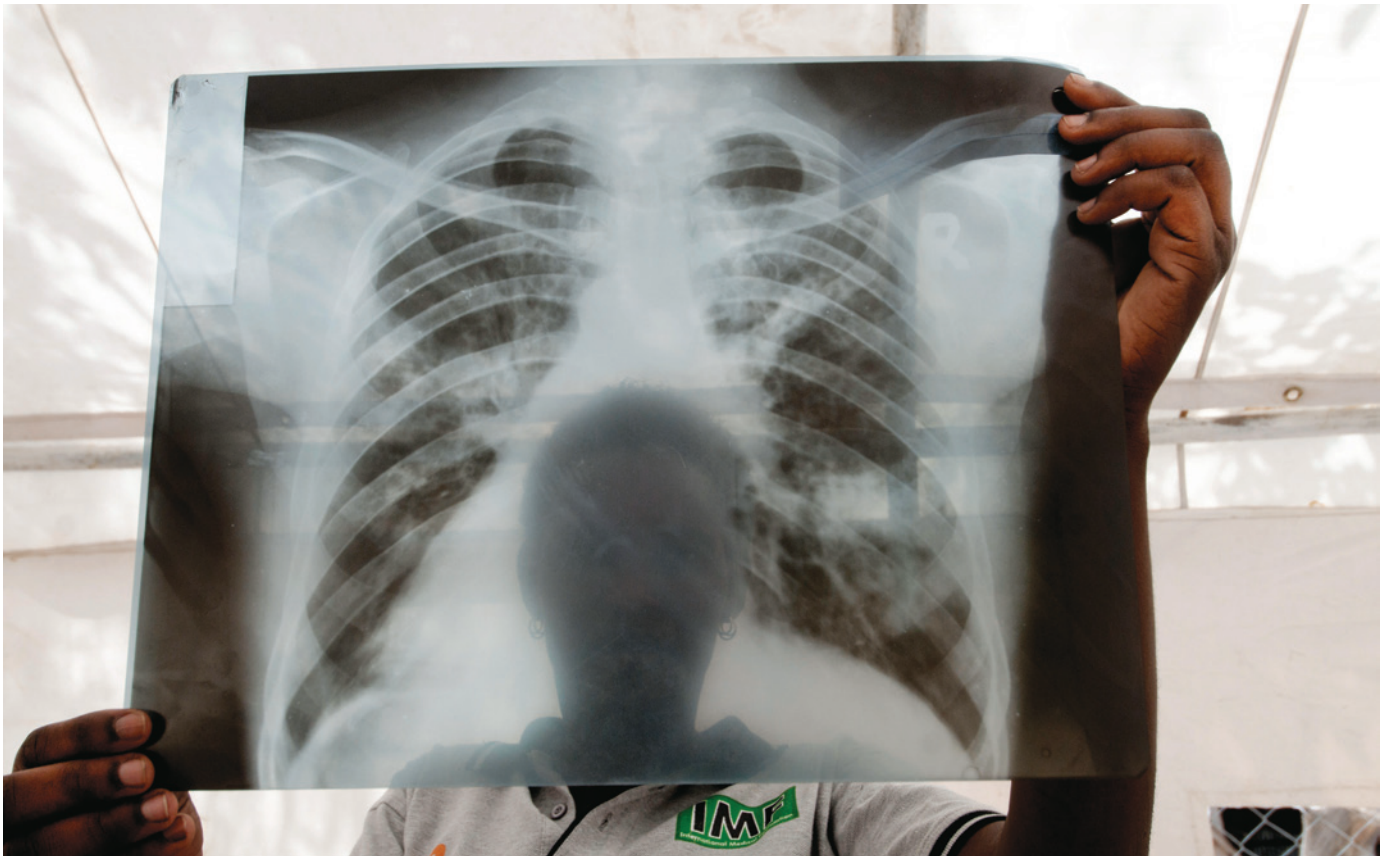
Maxine joined *Nature* in 1984, under the editorship of John Maddox, having been a researcher in the biophysics of muscle contraction. During her time at *Nature*, she held a variety of important posts before assuming her final role as Publishing Executive Editor.

GEORGE MARSHALL

Throughout her career she displayed a passionate commitment to *Nature*, focusing particularly on its scientific standards, its championship of intellectual integrity, the quality of its text, visual presentation, production standards and workflows, and the management of its staff. She never ceased to care and advocate for the working lives of those whom she managed. But, as I know from unsolicited compliments from authors over the years, she was highly valued outside the office too.

As a senior colleague put it: "Maxine was *Nature* through and through." Having been thankful for her many insights over the years, my colleagues and I will miss her greatly. ■

Philip Campbell, *Editor-in-Chief,* Nature.



J. MATTHEWS/PANOS

A chest X-ray from a patient with tuberculosis (TB) in Lira, Uganda. Uganda is one of 22 countries accounting for roughly 80% of new TB cases each year.

TB'S REVENGE

BY LEIGH PHILLIPS

The world is starting to win the war against tuberculosis, but drug-resistant forms pose a new threat.

If there was any doubt that tuberculosis (TB) was fighting back, it was dispelled in 2005, at the Church of Scotland Hospital in the village of Tugela Ferry, South Africa. Doctors at the hospital, in a rough, remote corner of KwaZulu-Natal province, were hardened to people dying from gunshots and AIDS. But even they were puzzled and frightened when patients with HIV who were responding well to antiretroviral drugs began dying — rapidly — from TB.

With ordinary TB, patients start to feel better after a few weeks or months on a selection of four mainstay antibiotics. But of the 542 people with TB at the hospital in 2005 and early 2006, 221 (41%) had a multidrug-resistant (MDR) form, against which these therapies are mostly powerless. Worse, 53 of them did not even respond to the few antibiotics that form a second line of defence. Eventually, doctors had nothing left to try: all but one of the 53 died, half of them within 16 days of diagnosis. It was the first major outbreak of what became known as extensively drug-resistant (XDR) TB — and a wake-up call to the world that TB had taken a turn for the worse[1].

In the early 1980s, TB cases had dropped to such low rates that Western policy-makers frequently talked of eradication of the disease. Then came the HIV epidemic, which triggered a resurgence of TB in the late 1990s. But the latest report on TB from the World Health Organization (WHO), published in October, revealed signs of progress against normal — or drug-sensitive — cases of the bacterial disease. New infections have fallen and the mortality rate has dropped by 41% since 1990. But, the

report warned, "drug-resistant TB threatens global TB control". Some 3.7% of new cases and 20% of previously treated cases are MDR-TB. And whereas in 2000 the highest incidence of MDR-TB was 14%, in Estonia; in 2010 that figure had jumped to 35%, in Russia's Arkhangelsk province. An estimated 9% of drug-resistant cases are XDR-TB, which has now been reported in 84 countries.

It is a tale of two TBs. Once detected, drug-sensitive TB is almost always treatable, as long as the appropriate drugs are provided and taken. Simple practices — such as checking that patients take their medicine — can be transformative. But in some countries, particularly in eastern Europe, Asia and Africa, the weakening or collapse of health-care systems over the past two decades has meant that patients do not always finish their drugs, or they take the wrong ones, allowing highly transmissible, drug-resistant strains to emerge and spread.

Drug-resistant TB is harder, more expensive and more time-consuming to treat. New tools are needed — but there have been no new anti-TB drugs in more than 50 years, and the current vaccine is largely ineffective. The most common diagnostic technique — analysing sputum samples under a microscope — can determine that *Mycobacterium tuberculosis* bacteria are present but not whether they are drug resistant. Meanwhile, researchers have lacked interest in developing drugs and tests, and drug companies have lacked market incentives to do so.

The growth of multi-drug resistance is an "escalating public-health emergency", says Grania Brigden, TB adviser for Médecins Sans Frontières (Doctors Without Borders) in Geneva, Switzerland: "With barely 1 in 20 TB patients being tested for drug resistance, we're just seeing the tip of the iceberg."

But scientists are careful to temper their alarm. In the past decade, researchers and policy-makers have fought for and won a reversal in funding and attention for TB. Several new drugs are in development, and progress is being made towards an effective vaccine.

"I do worry when people stand up at conferences and talk about MDR-TB and say it's a big disaster and the whole world is going to collapse. It's not that severe yet," says Tim McHugh, head of the Centre for Clinical Microbiology at University College London, who leads a team that is trialling one of the two most advanced candidates for new TB drugs. "The big anxiety is that if we don't act now, it will easily run away from us."

RISE AND FALL

TB is one of the world's leading killers, stealing 1.4 million lives and causing 8.7 million new and relapse infections in 2011. One-third of the world's population carries the bacterium, but most will never develop the active form of the disease.

The first modern TB epidemic took off in the late 1700s, during the Industrial Revolution. Rural workers in Europe and North America moved in droves to cities, where poverty — and related malnutrition and overcrowding — created an ideal environment for the disease's spread. But as hygiene, nutrition and medicine improved, what was known as the Great White Plague began to ebb.

"By the 1940 and 50s, things looked quite bright," says McHugh, who seems almost as interested in the history of TB as in its microbiology. The Bacillus Calmette–Guérin (BCG) vaccine, first used in the 1920s, helped. But BCG is now effective mainly against childhood TB, which is not infectious, rather than the adult form. What really broke TB's back was the introduction of isoniazid, in 1952, and then rifampicin, in the 1970s. "If you look at a graph of TB from the 1950s onward, [infection rates] just collapsed," McHugh says.

Then, in the 1980s and 1990s, HIV hit. "You can't underestimate the importance of HIV," McHugh says. A co-infection of TB and HIV produces a powerful biological synergy, accelerating the breakdown of the body's immune defences; latent TB infection is 20–30 times more likely to become active in people who have HIV. In 1993, the WHO declared TB a global emergency. Worldwide, TB is now the leading cause of death among people with HIV.

The resurgence of ordinary TB set the stage for drug-resistant forms. Resistance develops when people do not stick to their drug regimens — which typically last six months for drug-sensitive TB and 20 months for MDR-TB — allowing naturally occurring resistant mutants to grow and evolve. MDR-TB, which grew more threatening during the 1990s, is resistant to isoniazid and rifampicin. People with this form require second-line drugs — broad-spectrum antibiotics called fluoroquinolones or injectable agents (amikacin, capreomycin and kanamycin). These treatments are less effective, more toxic and take many more months to work than first-line therapies. An infection is classified as XDR-TB if it is also resistant to fluoroquinolones and at least one of the injectables. The XDR-TB outbreak at Tugela Ferry, when it was reported in 2006, rattled the TB research and policy community.

Experts agree that the biggest driver for the growth in drug-resistant TB has been the deterioration in some countries' health-care infrastructures, including TB programmes, since the 1990s — particularly in the former Soviet bloc. This decline has meant that patients are not diagnosed and treated; and in some countries, over-the-counter availability of anti-TB drugs also encourages people to take inappropriate second-line therapies, accelerating the growth of drug resistance. In a stroke of bad luck, the virulent and often drug-resistant 'Beijing' strain of TB, identified in 1995 in China, swept through Russia and eastern Europe just as the region's public-health provision was being dismantled. "There was a confluence of the biology of the organism and its progress into Russia, where lots of people had a health-care system that was collapsing around their ears," says McHugh. (In 2010, the Beijing strain was found in around 13% of active TB infections worldwide.) All of this helps to explain why the recent WHO report shows the highest burden of MDR TB in Russia's Arkhangelsk province and in Belarus, Estonia, Kazakhstan, Kirghizia and Moldova (see 'Two faces of TB').

FIGHTING BACK

Over the past decade or so, the paths of drug-sensitive and drug-resistant TBs have diverged. The solution for drug-sensitive TB is simply to deliver the drugs and diagnostics to patients — and the drive to do so has grown. One of the United Nations Millennium Development Goals set in 2000 was to halt and begin to reverse the incidence of TB by 2015; in 2001, the international Stop TB Partnership was established, bringing together government programmes, researchers, charitable foundations, non-governmental organizations (NGOs) and the private sector.

One major result of these and other efforts was a global expansion of 'directly observed treatment, short course' (DOTS), a strategy promoted by the WHO to combat drug-sensitive TB. Once diagnosed, the disease is treated with a supply of first-line drugs, taken under the close observation of health-care workers to ensure that people finish the course. Thanks in large part to such efforts, the WHO says that the world is on track to halve TB mortality from 1990 levels by 2015.

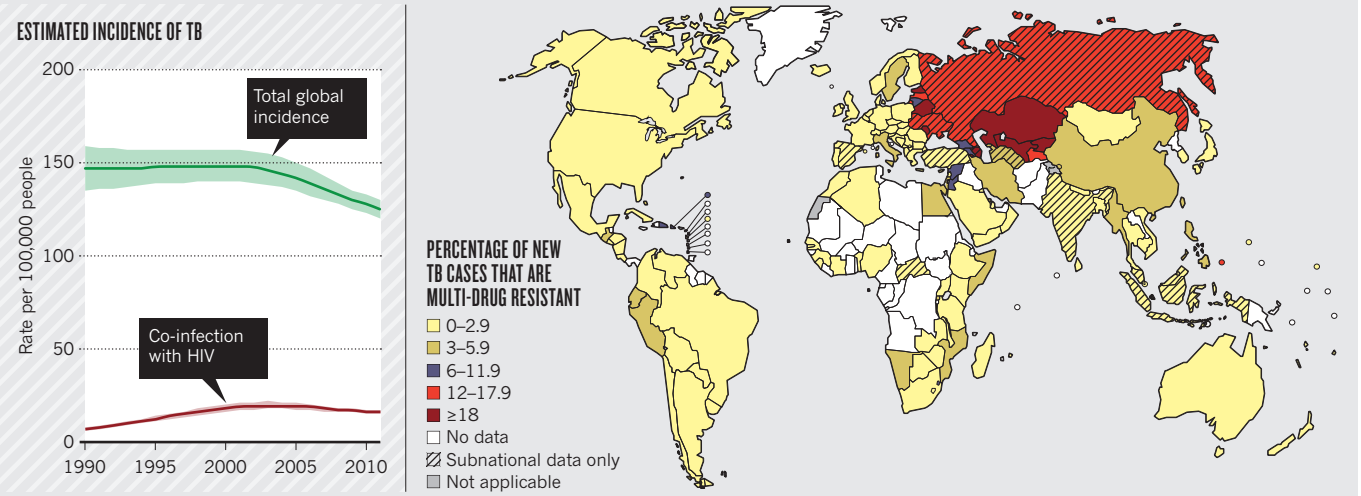
"WITH BARELY 1 IN 20 PATIENTS WITH TB BEING TESTED FOR DRUG RESISTANCE, WE'RE JUST SEEING THE TIP OF THE ICEBERG."

Tackling drug-resistant TB, however, will require not just the rebuilding of health-care infrastructure, but also new weapons, such as diagnostics, drugs and vaccines. The private sector has had little incentive to invest in basic research whose eventual products, if any emerge, would largely be sold at low cost in poor countries. "They have to look at the bottom line," says Anthony Fauci, head of the US National Institute of Allergy and Infectious Diseases in Bethesda, Maryland.

In the past decade or so, global TB programmes have pumped money into research. One major development came in 1998, when researchers at the Wellcome Trust Sanger Institute in Hinxton, UK, published the genome sequence of *M. tuberculosis*,

TWO FACES OF TB

Efforts to tackle drug-susceptible strains have started to cut the global incidence of TB (left), but collapsing health-care systems in the former Soviet bloc have helped drug-resistant strains to emerge and spread.



SOURCE: WHO

allowing researchers to identify and study genes that underlie the bacterium's virulence and ability to evade the immune system[2]. In 2012, the US National Institutes of Health in Bethesda started a bigger genome-sequencing project that aims to uncover the genetic roots of drug resistance. "We'll use next-generation sequencing technologies to sequence 1,000 TB clinical isolates from around the world — South Africa, Korea, Russia, Uganda — anywhere drug-resistant TB is heavily present," says Fauci.

WAR CHEST

There are now ten TB drugs in clinical trials. The aim is to find compounds that are effective against resistant strains and that work faster and have fewer side effects, so that patients will be more likely to finish the course. McHugh and his team, for example, are running a clinical trial at sites across Africa and Asia to test the antibiotic moxifloxacin, which is commonly used for pneumonia and skin infections. (They expect to release preliminary results in 2013.) The researchers are also working to speed up the screening process for potential drugs by using mycobacterial species that are less pathogenic and more fecund than *M. tuberculosis*, which is slow-growing, finicky and poses a biosecurity risk. "Previously, what you had was a chemist who says 'I've got this molecule that will kill *Escherichia coli*. I'm fairly sure it should kill TB. But there's nowhere I can see if it does,'" McHugh says.

Accurate and fast diagnostic tests for drug-resistant strains are also a key part of the fight, and a number of tests have come online in the past five years. One, called GeneXpert, takes 90 minutes to complete and is based on a gene-amplification technique that detects DNA sequences specific to *M. tuberculosis* and to rifampicin resistance. The system has been endorsed by the WHO and subsidized by a coalition of organizations, but researchers are still seeking simpler, cheaper options.

Only better vaccines will solve the problem for good. "We must invest in vaccine research if our ultimate goal is to be able to prevent the disease rather than forever chase growing drug resistance," says Helen McShane, a vaccine researcher at the University of Oxford, UK.

In 2008, the European Commission pushed for the creation of the TB Vaccine Initiative, which draws funding from European countries, NGOs and private funders. These and other efforts have helped to boost the number of vaccine candidates from 0 to 12 since 2000.

McShane and her team are on the cusp of the first efficacy results for MVA85A, one of the most clinically advanced TB vaccines in the pipeline at present. The shot, which McShane helped to develop as a PhD student 15 years ago, contains a virus designed to ramp up the activity of T cells that have already been primed by BCG. In 2009, in partnership with the South African Tuberculosis Vaccine Initiative, McShane launched a major phase II clinical trial on nearly 3,000 BCG-vaccinated babies in South Africa; early results are expected in the first quarter of 2013. In parallel, she and her colleagues are also testing the vaccine's efficacy in HIV-infected adults in South Africa and Senegal.

> *"WE MUST INVEST IN VACCINE RESEARCH IF OUR ULTIMATE GOAL IS TO BE ABLE TO PREVENT THE DISEASE."*

But are these efforts enough? "Unfortunately not," concludes Karin Weyer, coordinator of laboratories, diagnostics and drug resistance at the WHO Stop TB Department in Geneva. Annual funding for TB diagnosis and treatment is expected to reach some US$4.8 billion in 2013 — but TB care and control are expected to demand up to $8 billion a year by 2015. The $600 million contributed to TB research in 2010 also falls well short of the $2 billion the WHO estimates will be needed annually — and the economic crisis has slowed financing across the board. "I need to be and want to be optimistic," says Weyer. "But we're still working with shoestring budgets compared to HIV."

Meanwhile, the bacterium is not resting. In December last year, clinicians in Mumbai, India, reported[3] the identification of 12 patients with what they termed totally drug-resistant TB, or TDR-TB. Similar claims had been made a few years earlier in Italy and Iran, but this time the WHO took it seriously enough to investigate. In March 2012, 40 experts convened by the WHO concluded that there was not enough evidence to say that TDR-TB was substantially different from XDR-TB.

McHugh agrees. But he does not need further evidence to act. In the face of marching drug resistance, it is the responsibility of researchers to speak out, he says. "I think we can no longer be scientists in our labs doing fascinating stuff and think we're doing good work. We have to evangelize a little bit too." ■

Leigh Phillips *was an International Development Research Council fellow with* Nature *until October 2012.*

1. Gandhi, N. R. *et al. Lancet* **368,** 1575–1580 (2006).
2. Cole, S. T. *et al. Nature* **393,** 537–544 (1998).
3. Udwadia, Z. F., Amale, R. A., Ajbani, K. K. & Rodrigues, C. *Clin. Infect. Dis.* **54,** 579–581 (2012).

COMMENT

ARCHIVING Researchers must preserve their work for future historians of science **p.19**

ARTS Cultural highlights for 2013, from Laika opera to genome show **p.22**

INNOVATION The visionaries bent on transcending human limits **p.24**

CENTENARY Prize marks 100 years of the journal *Naturwissenschaften* **p.26**

WAQAR HUSSEIN/EPA/CORBIS

Flooding in Pakistan in September 2012 affected millions of people, displacing them and damaging their homes, farms and supplies of food and water.

Improve weather forecasts for the developing world

Global prediction partnerships would cost little and reduce the regional carnage caused by floods, droughts and tropical cyclones, argues **Peter J. Webster**.

Hurricane Sandy hit the northeast coast of the United States in October with ample warning. The storm caused widespread damage, but only around 125 people died in the region, thanks to planning made possible by accurate long-range weather forecasts.

In the developing world, tropical cyclones, floods and droughts arrive with little notice and kill thousands of people each year. Although only 5% of tropical cyclones occur in the north Indian Ocean, they account for 95% of such casualties worldwide[1,2]. In 2007 and 2008, two Very Severe Cyclonic Storms — Sidr and Nargis — caused the deaths of more than 10,000 and around 138,000 people in Bangladesh and Myanmar, respectively.

Flooding in the Ganges and Brahmaputra river basins has displaced more than 40 million people in each of the past few years[3,4]. In 1998, 60% of Bangladesh was inundated with floodwater for almost three months. Pakistan's Indus Valley was devastated by floods in 2010, costing more than 2,000 lives and US$40 billion[5]. Flooding struck the area again in the summers of 2011 and 2012.

Droughts condemn millions to hunger across the developing world. A three-week break in rainfall just after seasonal planting, following what seemed to be a normal monsoon onset, caused a disastrous crop failure in India in 2002 (ref. 6). The unforeseen dry period was tied to a south Asian weather oscillation that occurs every 30–60 days[7].

An individual living in south Asia or Africa can expect to encounter several extreme weather events in his or her lifetime. Because the resilience of poor populations is low and falls with every crisis, the cumulative effects are relentlessly impoverishing. Smallholders often purchase stocks on credit that is repaid at the end of the season, so the loss of a crop or livestock in one bad year can put the farmer into debt for many years, ▶

condemning generations to poverty.

Owing to advances in prediction science, such catastrophes can be forecast anywhere in the world with as long a lead time as Hurricane Sandy. The problem is tailoring complex global forecasts to a country or region and communicating them to local populations so that they can take action.

To lessen the impacts of adverse weather, networks must be established between the forecasters of global weather and climate in the developed world, and research, governmental and non-governmental organizations in the less-developed world. The investment needed is not high and will pay for itself: my research group at the Georgia Institute of Technology estimates that such a network could forecast floods across Asia for as little as $1 million a year, saving billions of dollars and thousands of lives.

NO WARNING

In most developing countries, weather warnings are issued a few days in advance, if at all. Yet, for a flood or cyclone, at least a week of forewarning is needed to allow the slowest members of a society (perhaps a farmer and his cattle) to evacuate. For short droughts, several weeks' notice allow smallholders to adjust planting and harvesting schedules. Long droughts require warnings to be given months ahead, so that farmers can choose resistant crops and store fodder and water.

Although Pakistan's devastating 2010 floods arrived with no warning in the north of the country, the pulses of intense rain responsible could have been forecast 8–10 days ahead if available data had been analysed at the time[5]. In September 2012, my research group used rainfall forecasts and a regional hydrological model to predict flooding ten days in advance, but Pakistan's government didn't implement the warning. Fortunately, the floods were less severe than those in 2010.

Regional forecasts with long lead times must take global atmospheric circulations into account, because local weather is influenced by distant events. The path of Hurricane Sandy, for example, was swayed by a mid-latitude low-pressure system in the Pacific Ocean that moved slowly across Canada, thousands of kilometres to the west.

Global weather-forecast models take decades to develop, are expensive to build and maintain, and are run by only a few national or multinational government organizations, including the European Centre for Medium-Range Weather Forecasts (ECMWF), the UK Met Office and the US National Centers for Environmental Prediction (NCEP).

Each run of a model begins with a huge array of meteorological and oceanographic data from more than 30,000 observations on land, 3,500 floating buoys and drifters, and numerous satellites. Even so, the global data coverage is incomplete (particularly in the Southern Hemisphere), the measurements have errors and the algorithms are imperfect. Global models are thus run many times a day using different initial conditions[8]. The ECMWF, for example, runs its model 51 times twice a day, incorporating new initial data to produce 1–15-day forecasts, and extends its forecasting horizon to 32 days ahead twice a week.

In theory, developing countries can access these data streams. The NCEP forecasts, for example, are posted on the Internet daily. But peeling off regional data from the global deluge is like filling a cup from a fire hose. Internet transmission costs are high, and timely downloading requires a fast data-transfer rate. Less-developed countries have small budgets and slow Internet connections. Paradoxically, as forecasts become better and their resolution grows, it becomes more difficult for developing countries to access them.

Regional forecasting requires calibration with local data, such as geographic contours, and so is beyond the remit of the global weather centres. Intermediaries — research groups, universities or companies — can form a bridge between global-forecast providers and user communities. Partnerships are springing up, but more are needed.

Bangladesh offers a success story that could be emulated elsewhere. Global forecasts produced in Europe are sent to the United States and turned into flood forecasts that, within six hours, can be integrated into Bangladesh's disaster-management protocol by local experts.

The need for a rapid forecasting and warning system in Bangladesh became apparent following the 1998 floods. The ECMWF, the Bangladeshi government and my research group therefore developed a 1–10-day flood-forecasting system and created the Climate Forecast Applications Network (CFAN) to distribute it. This system was first used experimentally in 2004 and became operational in 2007. The basic science was developed with support from the US National Science Foundation, and implementation by the CFAN was funded by the US Agency for International Development (USAID) and the humanitarian agency CARE.

The CFAN has produced daily forecasts of the Brahmaputra and Ganges flows since 2004, sending them to the Bangladesh Flood Forecast and Warning Centre[3,4]. If the probability of flooding exceeds 80%, warnings are issued to government offices across Bangladesh.

Planning and training are essential for effective use of the forecasts. Before the 2007 flood season, village and community leaders in six administrative unions of Bangladesh were trained to interpret the data and to take action if flooding was likely. Local leaders could tell farmers to harvest crops, shelter animals, store clean water and secure food, household and farming effects.

Bangladesh experienced three major floods in 2007 and 2008. Each was forecast successfully ten days in advance and mitigation steps were taken[3,4] (see 'Bangladesh flood alerts'). On the basis of a World Bank report[9], one analysis concluded that about $40 was saved for every dollar invested in the regional forecasting and warning system. Savings at the village level were measured in units of annual income[4].

In 2009, to boost regional capacity-building, the CFAN handed over its flood-forecast modules to the Bangladesh Flood Forecasting and Warning Centre. When the large volume of data proved too difficult for the centre to handle, the responsibility shifted to an international non-government entity, the Regional Integrated and Multi-Hazard Early Warning System (RIMES).

Funded partly by contributions from member states, RIMES works with governments across south and east Asia to incorporate regional forecasts into national disaster-mitigation programmes and provides warnings for a range of natural hazards, including

BANGLADESH FLOOD ALERTS

Bangladesh's floods of July–August 2007 were predicted ten days in advance. Forecasts were communicated to five pilot areas (red), where community leaders directed people to harvest crops, lead cattle to safety and store water, food and personal belongings.

SOURCE: CEGIS/UNOSAT

earthquakes, tsunamis and extreme weather. RIMES is innovative but its funding is limited, making it difficult to maintain the cadre of scientists necessary to tackle specific problems. Inadequate funding also hinders the essential updating of forecast modules as satellite and global forecast systems change.

GLOBAL PARTNERSHIPS

Partnerships that bridge the gap between the global forecasters and the user community need to be established in other regions to address a range of weather hazards. The plan and type of group that forms the bridge will depend on the type of hazard being addressed. But the aim of each team is the same: to produce hazard-forecast modules based on the global forecasts and to use them to provide warnings for the region. The team will also be responsible for updating the modules as systems and technologies change.

Such partnerships can be aided by sustained funding from intergovernmental organizations, such as the United Nations, the World Bank and USAID. My research group estimates that the cost of extended 10–15-day forecasts for south and east Asia for a wide range of hydrometeorological hazards (including slow-rise monsoon floods, droughts and tropical cyclones) is relatively small: perhaps $2 million to $3 million per year.

Asia and Africa stand on the threshold of great economic advancement and can build resilience through the effective use of longer-range weather forecasts[10]. Faced with possible climate change, societies that learn to cope with and mitigate hazards now will be most adept at dealing with more frequent and intense hazards in the future. ■

Peter J. Webster *is a professor of Earth and Atmospheric Sciences at the Georgia Institute of Technology, Atlanta, USA.*
e-mail: pjw@eas.gatech.edu

1. Webster, P. J. *Nature Geosci.* **1,** 488–490 (2008).
2. Belanger, J. I., Webster, P. J., Curry, J. A. & Jelinek, M. T. *Weather Forecast.* **27,** 757–769 (2012).
3. Hopson, T. M. & Webster, P. J. *J. Hydrometeorol.* **11,** 618–641 (2010).
4. Webster, P. J. *et al. Bull. Am. Meteorol. Soc.* **91,** 1493–1514 (2010).
5. Webster P. J., Toma, V. E. & Kim, H.-M. *Geophys. Res. Lett.* **38,** L04806 (2011).
6. Webster, P. J. & Hoyos, C. *Bull. Am. Meteorol. Soc.* **85,** 1745–1765 (2004).
7. Stephens, G. L. et al. *J. Clim.* **17,** 2213–2224 (2004).
8. Leutbecher, M. & Palmer, T. N. *J. Comp. Phys.* **227,** 3515–3539 (2008).
9. Teisberg, T. J. & Weiher, R. F. *Background Paper on the Benefits and Costs of Early Warning Systems for Major Natural Hazards* (World Bank, 2009).
10. Foresight *Reducing the Risks of Future Disasters: Priorities for Decision Makers* (UK Govt Office of Science, 2012).

M. ARGLES/GUARDIAN NEWS & MEDIA

The correspondence of Francis Collins (left) and John Sulston illuminates a vital part of science history.

Science today, history tomorrow

We must preserve the interactions of contemporary researchers for future scholars, urges **Georgina Ferry**.

The year 1998 was crucial for the Human Genome Project (HGP), an international collaboration launched eight years before to sequence the complete human genome. Spurred by the launch of a privately financed sequencing bid by Craig Venter, the HGP's leaders decided to accelerate their own efforts. Some of the proposed changes caused friction — the HGP was long planned and carefully executed. In October that year, John Sulston, then director of the Wellcome Trust Sanger Institute near Cambridge, UK, felt so beleaguered that he sent a strongly worded e-mail to Francis Collins, then director of the US National Human Genome Research Institute (NHGRI) in Bethesda, Maryland. The subject line? 'Friendly fire'. ▶

earthquakes, tsunamis and extreme weather. RIMES is innovative but its funding is limited, making it difficult to maintain the cadre of scientists necessary to tackle specific problems. Inadequate funding also hinders the essential updating of forecast modules as satellite and global forecast systems change.

GLOBAL PARTNERSHIPS

Partnerships that bridge the gap between the global forecasters and the user community need to be established in other regions to address a range of weather hazards. The plan and type of group that forms the bridge will depend on the type of hazard being addressed. But the aim of each team is the same: to produce hazard-forecast modules based on the global forecasts and to use them to provide warnings for the region. The team will also be responsible for updating the modules as systems and technologies change.

Such partnerships can be aided by sustained funding from intergovernmental organizations, such as the United Nations, the World Bank and USAID. My research group estimates that the cost of extended 10–15-day forecasts for south and east Asia for a wide range of hydrometeorological hazards (including slow-rise monsoon floods, droughts and tropical cyclones) is relatively small: perhaps $2 million to $3 million per year.

Asia and Africa stand on the threshold of great economic advancement and can build resilience through the effective use of longer-range weather forecasts[10]. Faced with possible climate change, societies that learn to cope with and mitigate hazards now will be most adept at dealing with more frequent and intense hazards in the future. ■

Peter J. Webster *is a professor of Earth and Atmospheric Sciences at the Georgia Institute of Technology, Atlanta, USA.*
e-mail: pjw@eas.gatech.edu

1. Webster, P. J. *Nature Geosci.* **1,** 488–490 (2008).
2. Belanger, J. I., Webster, P. J., Curry, J. A. & Jelinek, M. T. *Weather Forecast.* **27,** 757–769 (2012).
3. Hopson, T. M. & Webster, P. J. *J. Hydrometeorol.* **11,** 618–641 (2010).
4. Webster, P. J. *et al. Bull. Am. Meteorol. Soc.* **91,** 1493–1514 (2010).
5. Webster P. J., Toma, V. E. & Kim, H.-M. *Geophys. Res. Lett.* **38,** L04806 (2011).
6. Webster, P. J. & Hoyos, C. *Bull. Am. Meteorol. Soc.* **85,** 1745–1765 (2004).
7. Stephens, G. L. et al. *J. Clim.* **17,** 2213–2224 (2004).
8. Leutbecher, M. & Palmer, T. N. *J. Comp. Phys.* **227,** 3515–3539 (2008).
9. Teisberg, T. J. & Weiher, R. F. *Background Paper on the Benefits and Costs of Early Warning Systems for Major Natural Hazards* (World Bank, 2009).
10. Foresight *Reducing the Risks of Future Disasters: Priorities for Decision Makers* (UK Govt Office of Science, 2012).

M. ARGLES/GUARDIAN NEWS & MEDIA

The correspondence of Francis Collins (left) and John Sulston illuminates a vital part of science history.

Science today, history tomorrow

We must preserve the interactions of contemporary researchers for future scholars, urges **Georgina Ferry**.

The year 1998 was crucial for the Human Genome Project (HGP), an international collaboration launched eight years before to sequence the complete human genome. Spurred by the launch of a privately financed sequencing bid by Craig Venter, the HGP's leaders decided to accelerate their own efforts. Some of the proposed changes caused friction — the HGP was long planned and carefully executed. In October that year, John Sulston, then director of the Wellcome Trust Sanger Institute near Cambridge, UK, felt so beleaguered that he sent a strongly worded e-mail to Francis Collins, then director of the US National Human Genome Research Institute (NHGRI) in Bethesda, Maryland. The subject line? 'Friendly fire'. ▶

For anyone interested in the history of the HGP, this e-mail is a key document (and one that was later acknowledged as an 'emotional outburst' by its author, who now leads the Institute for Science, Ethics and Innovation at the University of Manchester, UK). Was it a catalyst for improved communication between the main players? Who helped to resolve the conflict? To what extent were the directors of the five leading sequencing centres competing as well as collaborating? The content of the e-mail traffic between and within the sequencing teams offers a potentially rich seam of enquiry.

As the co-author of Sulston's account of the HGP (*The Common Thread*; Joseph Henry Press, 2002), I saw this e-mail and many others, but they are not generally available. That may change, thanks to an international archiving programme now under way. The Wellcome Library in London is funding an archivist, Jenny Shaw, to survey the documentary record relating to the HGP and earlier mapping and sequencing activities in the United Kingdom between 1977 and 2004. Ludmila Pollock, executive director of the library and archives at Cold Spring Harbor Laboratory (CSHL) in New York, is conducting a parallel exercise in the United States. The first objective is to catalogue these materials. A longer-term aim — which will depend heavily on funding and the willingness of the scientific community to cooperate — is to secure them in reputable repositories and make them available to scholars.

The programme throws into relief how fragile a trace modern science is leaving in the historical record. As a scientific biographer, I have spent hours happily immersed in piles of yellowing papers that are carefully stored in archive boxes and guarded by watchful custodians in academic libraries. Future biographers will not be as lucky. Today's scientists underestimate the historical importance of anything other than their published papers; they communicate almost entirely electronically; and funding for archival preservation is increasingly uncertain. If we care about documenting the astonishing discoveries of the twentieth and twenty-first centuries, we must act now.

GOOD PRACTICE

Why are archiving exercises such as the HGP's necessary? We are fortunate that Collins, who is now director of the US National Institutes of Health (NIH), kept his papers, and even more fortunate that his successor at the NHGRI, Eric Green, is employing an archivist to digitize them. The NHGRI is only now developing an archiving policy. Previously, says Green, 'records administration' at the institute meant throwing things away that were 'no longer needed'. But this material represents only a fraction of the documents that record the HGP's history. The genome-sequencing story began before the NIH took a leading role, and it involved many institutions and individuals — inside and outside the United States — for which the NIH had no responsibility.

CERN/SPL

Particle-accelerator building at CERN, which launched its archiving programme in 1979.

There remains an urgent need to reconstruct the 'paper trails' — often largely electronic — that eventually led to the complete, publicly available sequence. These include a vast amount of e-mail correspondence, as well as informal literature such as *The Worm Breeder's Gazette* (now exclusively online) and the software that Sulston and his colleague Richard Durbin wrote to manage their genome-mapping data in the mid-1980s.

"It is difficult to convince people that their tweets and IMs are the stuff of history."

Comprehensive record-keeping is easier within a single institution. CERN, the particle-physics laboratory near Geneva, Switzerland, showed a commendable sense of its own historical significance when it commissioned a regularly updated biography in 1979, 25 years after it opened. Divisional records officers at the facility now ensure a smooth pipeline through to the central archives. The archivists encourage senior scientists to have their filing systems appraised for historical interest, and they have a strategy for selecting and preserving e-mails. As the birthplace of the World Wide Web, CERN is also working to archive its own web pages.

Scientific archives typically consist of institutional records such as CERN's, and the personal papers of distinguished (and, usually, dead) scientists. When I wrote a biography of the Nobel prizewinner Dorothy

Hodgkin (1910–94), I relied heavily on her collected papers, housed in the Bodleian Library at the University of Oxford, UK. Among them I found, for example, a letter dated June 1939, in which Hodgkin's contemporary Dorothy Wrinch attempted — unsuccessfully — to win sisterly solidarity for her erroneous theories on protein structure ("Our chromosome count of course does not tend to weaken the desire of others to attack us," she wrote).

FLEETING TWEETS

But it is increasingly difficult for such repositories to capture the full picture. Modern science involves experts in numerous disciplines, collaborating within and between institutions. Few will have reached the age or eminence at which scientists might typically think of depositing their papers. The lab notebooks kept by technicians and junior research staff are also important. Most of the contributors to the HGP are still very much alive, so archivists will need permission to scan the contents of their e-mail accounts on personal hard drives or institutional servers.

The ubiquity of 'born-digital' material, such as e-mails and web pages — and now tweets, Facebook comments, forum posts and instant messages (IMs), which many scientists use to share expertise, announce new publications or engage in policy debates — makes the quest to preserve even more urgent. The hardware and software used to generate and store digital material can become obsolete or be deleted or otherwise destroyed with ease, and it is not always clear who owns it. The British Library in London is tackling these issues with its Digital Lives project; it also leads the UK Web Archive collaboration, which stores pages of scholarly interest. Several online tools for archiving tweets by hashtag or username are available.

Technology is the easy bit, however. Much more difficult is convincing people that their off-the-cuff tweets and IMs are the stuff of history. E-mail users do not always file their personal and professional messages separately, and many are understandably wary of making any of their correspondence public. Since the 'Climategate' affair in 2009 — when e-mail servers at the University of East Anglia, UK, were hacked and their contents publicized — scientists are all too aware that an unguarded remark to a colleague can snowball into a cause célèbre.

But archivists have long experience in handling questions about confidentiality and access. The University of Cambridge library holds six boxes of love letters that belonged to the UK crystallographer J. D. Bernal, which are not to be opened until 2021 — 50 years after his death. I would trust any of the archivists I have worked with to respect my wishes.

"Let's not wait until memories have faded and papers been discarded … before deciding to save our heritage," exhorted Nobel prizewinner Sydney Brenner in 2007, in a letter announcing the donation of his papers to the CSHL archive (S. Brenner and R. J. Roberts *Nature* **446,** 725; 2007). Others may ask why historians have any right see material not in the published record. Sulston himself, who is supportive of the archiving project, admits that "scientists can be quite conflicted about historical research, because it's important to forget things and move on."

The history of science is much more than a chronology of scientific facts and theories. Access to informal sources is essential for understanding the personal, political and social context in which research takes place. Take the story of the discovery of the double-helical structure of DNA, told by historian Robert Olby in *The Path to the Double Helix* (Dover Publications,1974), by science writer Horace Freeland Judson in *The Eighth Day of Creation* (Simon and Schuster, 1979) and revisited more recently in the biographies of Rosalind Franklin and Francis Crick.

> ***"We want our scholarly successors to be able to follow the twists and turns of the scientific, political and personal pathways."***

All these authors had access to correspondence and notebooks, without which we would have been none the wiser about the complex web of personal communication (and miscommunication) that led Crick and James Watson to their discovery. And when more turned up in 2010, the plot thickened again. Two decades of Crick's correspondence turned out to be mixed up with Brenner's papers (the two shared an office from 1956 to 1977), including a 1953 letter in which Crick admitted to a colleague that if he had seen Rosalind Franklin's photo of the 'A' form of DNA, he would have been "considerably worried" (A. Gann and J. Witkowski *Nature* **467,** 519–524; 2010).

SAVE THE SAVERS

Fifty years hence, do we want the story of the sequencing of the human genome — an enterprise that forms the bedrock of much of twenty-first-century biomedicine — to rest on scientific papers and institutional reports, leavened only with breathless news reporting from the popular media? Surely not. We want our scholarly successors to be able to follow the twists and turns of the scientific, political and personal pathways that intersected as the human genome's 3 billion base pairs winked across the screens of the sequencing centres.

So, what to do? First, institutions and individuals need to have the confidence to place their records in the hands of professional archivists working in reputable repositories. Second, funding bodies need to be convinced that efforts to acquire, store, catalogue and disseminate such records are essential to preserving our scientific heritage.

In the United Kingdom, the funding situation is grim. The UK National Cataloguing Unit for the Archives of Contemporary Scientists had a 36-year track record of seeking out and cataloguing the papers of British scientists before losing its core funding and closing in 2009. A successor organization, the Centre for Scientific Archives (CSA), was established at a Science Museum store in Wroughton, UK, the same year. The CSA is run by a voluntary board chaired by Anne Barrett, an archivist at Imperial College London, and it receives some in-kind help from bodies such as the Royal Society and the National Archives.

The CSA has no core funding. At present, it engages freelance archivists to catalogue half a dozen collections of personal papers (including those of physicists Joseph Rotblat and Gareth Roberts) that came with their own funding. Meanwhile, the post of curator of the history of science at the British Library (which holds the papers of luminaries such as Charles Babbage and Alexander Fleming) has been frozen since it was vacated in February 2011.

Admittedly, history is not a priority at a time of recession. But it is baffling that scientific heritage has attracted so little support in the United Kingdom, especially as the country managed to raise almost £8 million (US$12.9 million) to keep an 1868 painting (*Portrait of Mademoiselle Claus*) by the French artist Edouard Manet from being moved overseas in August 2012. The project to catalogue the HGP is benefiting from the Wellcome Trust's long-standing commitment to the history of medical research, and in the United States, the CSHL has found early support from private donors and foundations.

But much more funding is needed. And it is urgent that we carry out similar scoping studies for other areas of science, to identify what should be saved before it disappears.

The sequencing of the human genome did not answer every question in biology, but it provided a publicly accessible resource that can be used for all time by inventive scientists with new questions. A documentary archive of the project will provide just such a resource for historians. If we want future generations to understand how science and society interact, libraries, research institutions and individual researchers must work together to preserve the documentary heritage of contemporary science. ■

Georgina Ferry *is a science writer and author and a member of the advisory committee of the Wellcome Library's Human Genome Archive Project, Oxford OX2 6JE, UK.*
e-mail: mgf@georginaferry.com

The opera *Laika the Spacedog* will tell the story of the first animal to be sent into orbit.

LISTINGS

Hot tickets for 2013 in science and art

This is your year if you want to rub shoulders with canine cosmonaut Laika or astronomer Galileo Galilei; travel through time, oscillate, get lost in a fog sculpture or ponder extinction; or listen to sound projected through liquid nitrogen. **Jascha Hoffman** offers his top tips on science's cultural calendar.

Laika the Spacedog
Science Museum, London
22–25 January (then on tour)

Canine cosmonaut Laika was the first animal launched into orbit, in 1957. This "highly interactive" opera about the Soviet pooch is intended to galvanize science-hungry schoolchildren. Physical forces, the Solar System, rocket science and health are explored through singing, animated film, puppetry and live action in this English Touring Opera production, with music by Russell Hepplewhite and lyrics by Tim Yealland. There is even a chance to play a theremin. A happy ending may not be in the offing, however: a victim of the space race between the Soviets and Americans, Laika died in orbit.

A Life of Galileo
Swan Theatre, Stratford-upon-Avon, UK
31 January to 30 March

Bertolt Brecht's masterly evocation of science at bay, *A Life of Galileo*, gets the Royal Shakespeare Company treatment in this new production. Current writer-in-residence Mark Ravenhill has penned a fresh translation of the seventeenth-century physicist–philosopher's grim battles with the Church over heliocentrism. Ian McDiarmid takes the title role; Roxana Silbert directs. Ravenhill, whose debut was *Shopping and Fucking* (1996), is unabashedly political, so the Marxist tones of Brecht's work may be writ large.

JUDE MUNDEN

MICHAEL HANNA

OSCILLATOR
Science Gallery, Dublin
7 February to 14 April

Circadian rhythms, resonating bridges, erupting geysers and volatile markets — Dublin's dynamic Science Gallery kicks off its 2013 season with an exploration of oscillation, including cultural fads that cycle in and out of vogue. Curator is media artist Douglas Repetto, founder of the worldwide network Dorkbot, which brings together artists, designers, scientists and innovators working in electronic art. Dizzy visitors may realize that everything comes in waves — even shows such as this one, which makes way for exhibitions on risk, illusion and synthetic biology later in the year.

Extinction: Not the End of the World?
Natural History Museum, London
8 February to 8 September

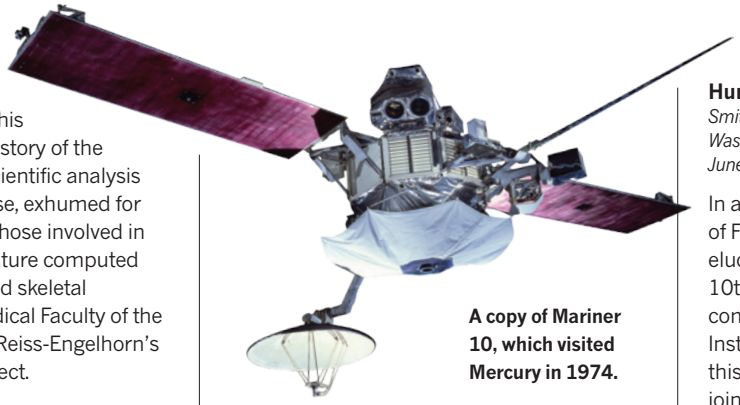
The main hall of London's iconic Natural History Museum is dominated by the 26-metre-long cast of a fossilized *Diplodocus* skeleton. Hundreds of the taxidermied beasts that throng its rooms, imprisoned in cabinets, are gone for good outside its walls. So the museum is the ideal spot to explore extinction, the subject of this major show. It will pull together images and specimens of species already lost, as well as those teetering on the brink, to probe issues around how human exploitation and other factors have depleted the variety of life.

NATURAL HISTORY MUSEUM

The Medici: People, Power and Passion
The Reiss-Engelhorn Museum, Mannheim, Germany
17 February to 28 July

Some of the most sumptuous art of the Italian Renaissance was bankrolled by the Medici — and the dynasty also supported science for centuries. Members of the illustrious clan backed the likes of Galileo Galilei and contributed to natural history, medicine and applied mathematics. When the last of the ducal line, Anna Maria Luisa de Medici, died in 1743, she bequeathed to the city of Florence the family treasures — and,

NATIONAL AIR AND SPACE MUSEUM, SMITHSONIAN INSTITUTION



A copy of Mariner 10, which visited Mercury in 1974.

unwittingly, a marvel of a different sort to posterity. This exhibition on the cultural history of the Medici will showcase the scientific analysis of Anna Maria Luisa's corpse, exhumed for the purpose in late 2012. Those involved in the project — which will feature computed tomography animations and skeletal remains — include the Medical Faculty of the University of Florence and Reiss-Engelhorn's own German-Mummy Project.

Time and Navigation: The Untold Story of Getting from Here to There
Smithsonian National Air and Space Museum, Washington DC
Opens 29 March

Timekeeping has been essential to navigation since the eighteenth century, when British carpenter John Harrison invented a clock accurate enough for sailors to determine their longitude using deviations in the positions of celestial bodies. And it is still a necessity in the age of the Global Positioning System; atomic clocks in orbiting satellites must be synced to within nanoseconds to allow mobile phones to triangulate their position. This exhibition, a collaboration between the Smithsonian Institution's National Air and Space Museum and National Museum of American History, takes in the history of navigation with an emphasis on timekeeping. It features the clock Charles Lindbergh used to navigate the *Spirit of St Louis* across the Atlantic in 1927; a spacecraft whose navigation system propelled it to Mercury in the 1970s using the gravity field of Venus; and a self-navigating Volkswagen named Stanley.

Margaret Guthman Musical Instrument Competition
Georgia Institute of Technology, Atlanta
11–12 April

This competition at Atlanta's public-research hotspot offers US$10,000 in prizes for innovations in "musicality, design and engineering". Winning entries in years past have included a device that turns everyday objects such as a whisk into musical instruments and a vintage slot machine by Berlin-based artist Christian Graupner and partners that allows players to remix musical video clips into an "audio-visual triptych". This year's judges (among them Berkeley computer musicologist David Wessel and performance artist Laurie Anderson) will face hard choices, if last year's inventions are anything to go by. Those included the textile-based *Audio Skin* and the *Resistor JelTone*, a part-edible toy piano by NYC Resistor, the New York-based hacker collective.

ZUM/ZUMLLC.COM

Exploratorium
New building at Pier 15, San Francisco, California
Opens 17 April

When it was opened in 1969 by physicist Frank Oppenheimer, who worked on the Manhattan Project with his better-known brother, Robert, the Exploratorium was one of the first museums to favour an interactive science education. Now, as it relocates to the city's waterfront, the Exploratorium is expanding its artist-in-residency programme through its new Center for Art & Inquiry. The museum will unveil site-specific works, including an Aeolian, or wind, harp by Doug Hollis that plays the breezes from the bay between Piers 15 and 17, and a 'fog sculpture' by Japanese artist Fujiko Nakaya.

Human Genome exhibition
Smithsonian National Museum of Natural History, Washington DC
June 2013 to June 2014

In a year that sees both the 60th anniversary of Francis Crick and James Watson's elucidation of DNA's structure and the 10th anniversary of the human genome's complete decoding, the Smithsonian Institution in is pulling out all the stops. For this exhibition, its natural history museum joins forces with the National Human Genome Research Institute in Bethesda, Maryland, to explore what the genome is, what it tells us and how this information could revolutionize health care and our understanding of our place in the world. After its time on the National Mall, the show will travel around North America.

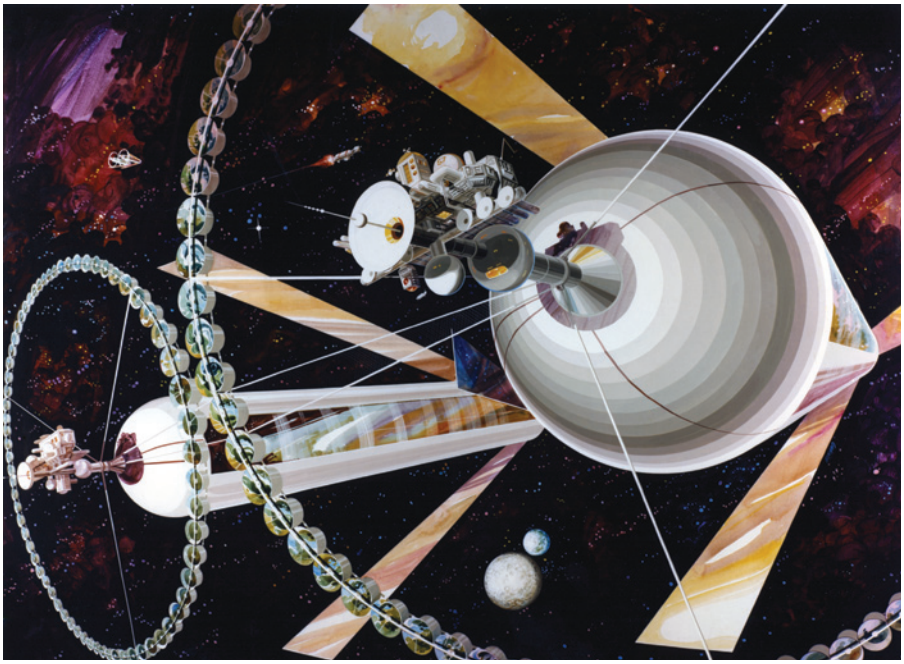
Ars Electronica Festival
Linz, Austria
Early September

Founded in Linz in 1979, Ars Electronica has long been the festival at the cutting edge of electronic art. Over the years it has expanded from galleries and conference halls to tunnels, factories and even a nearby monastery in its effort to bring technology-driven art to the public. This year, the festival teams up with the budding artist-in-residence programme at particle-physics powerhouse CERN, near Geneva in Switzerland, to premiere a new installation by US sound artist Bill Fontana. Mimicking the protocol of a scientific experiment, Fontana will project short bursts of sound through various substances — air, water, liquid nitrogen and metal, for example — to compare their speed of conduction and sonic properties.

Large Hadron Collider
Science Museum, London
8 November 2013 to 30 April 2014

After last year's thrilling announcement of the discovery of a Higgs boson, it was inevitable that the Large Hadron Collider, built and operated by some 10,000 scientists and engineers, would earn its own exhibition. When the travelling show debuts at London's Science Museum, visitors will see the small bottle of hydrogen gas that fed protons into the 27-kilometre accelerator at CERN, the world's largest particle-physics laboratory, near Geneva in Switzerland. Also on display will be a giant superconducting magnet used to bend the particle beams, which approach the speed of light, and historical devices such as the apparatus that led to the discovery of the electron in 1897.

Jascha Hoffman *is a journalist based in San Francisco, California. Additional reporting by Daniel Cressey and Alison Abbott.*



RICK GUIDICE/NASA

The Cylindrical Colony: one of several designs for a space settlement mooted by Gerard O'Neill.

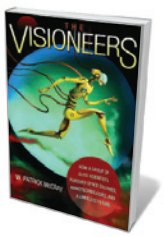
INNOVATION

Limits be damned

Cyrus Mody applauds an examination of the twentieth-century scientists who dreamed of breaking the bounds.

To the best of our knowledge, human life is constrained by natural limits: we do not live forever, we cannot transport ourselves or transmit information faster than the speed of light, and there is a finite supply of fossil fuels. Debates about such limits have shaped, and been shaped by, scientific and technological knowledge for centuries. Even faulty predictions about limits have made important contributions. Thomas Malthus' pessimism, for instance, prepared the ground for Darwin's theory of natural selection, and the overly optimistic vision of Lewis Strauss, former chairman of the US Atomic Energy Commission, of "energy too cheap to meter" facilitated decades of nuclear-power research and development.

In *The Visioneers*, science historian Patrick McCray of the University of California, Santa Barbara, argues that the resource-scarcity debates of the 1970s inspired a generation of visionary scientists and engineers. This influential crew had big dreams about overcoming all kinds of limits; occasionally built working models to demonstrate progress towards their dreams; and passionately assembled coalitions to make those dreams a reality.



The Visioneers: How a Group of Elite Scientists Pursued Space Colonies, Nanotechnologies, and a Limitless Future
W. PATRICK MCCRAY
Princeton University Press: 2012. 328 pp. £19.95, $29.95

McCray focuses on Gerard K. O'Neill, the Princeton physicist and designer of space colonies, and on his protégé, K. Eric Drexler, the 'speculative engineer' trained at the Massachusetts Institute of Technology (MIT) in Cambridge who helped to put nanotechnology on political agendas in the early 1990s. Along the way, McCray introduces a large and colourful cast of others who, over four decades, promoted technological progress as the way to overcome every limit.

O'Neill's ideas reached a mass audience in part through the L5 Society founded in 1975 by Keith and Carolyn Henson. These livestock farmers and Tolkien enthusiasts from Arizona later drifted into advocacy for the Strategic Defense Initiative and cryonic life extension, a proposed technology by which all or part of a human body would be frozen at death in the hope that it could be re-animated later. Drexler also imagined that cryonic immortality could be facilitated by programmable 'molecular assemblers' — nanometre-scale robots, or nanobots — repairing the tissues of corpses frozen at death.

Pillars of the California counterculture — such as the psychologist and LSD advocate Timothy Leary, and Stewart Brand, founder of the *Whole Earth Catalog* — also took up the visions of O'Neill and Drexler, working them into manifestos on transhumanism and the 'electronic frontier'. Brand served on the board of the Foresight Institute, set up in 1986 by Drexler, and made Drexler's molecular assemblers a centrepiece of the future scenarios that his Global Business Network sold to enthralled chief executives.

McCray documents how cryonics and radical life extension, space colonies, molecular nanotechnology and exotic sources of energy (such as solar-power satellites and zero-point energy) were widely popularized, alongside unsceptical articles about paranormal phenomena, by the pornographers Bob Guccione and Kathryn Keeton in their glossy monthly magazine *Omni*. Indeed, McCray argues that the audience that *Omni* catered to — young and male, with a taste for luxury goods, high-tech gadgets, libertarian politics and libertine excesses — strongly resembled the visioneers and many of their followers.

One thread ran through all of this: the 1972 blockbuster *The Limits to Growth* (Universe), by global think-tank the Club of Rome. This book goaded O'Neill and Drexler, says McCray, into sketching their plans for a limitless future. *The Limits to Growth* — along with public intellectuals such as the biologist Paul Ehrlich and the ecologist Garrett Hardin, plus fictional films such as *Soylent Green*, *Logan's Run* and *Silent Running* — popularized the idea that resource scarcity and a growing population would combine to create shortages of economically crucial materials. That message took root around the world in the 1970s, particularly (if temporarily) in a United States beset by 'stagflation', oil shortages and environmental crises such as the Santa Barbara oil spill of 1969.

However, the original computer models on which *The Limits to Growth* was based, developed by veterans of Jay Forrester's systems-dynamics group at MIT, failed to account adequately for the role of technological innovation in ameliorating resource scarcity, at least over the near term. Although the models were later refined, the 1972 version provoked a storm of criticism, much of it justified. Many lay people, particularly those of the generation whose childhoods

➲ NATURE.COM
Hugh Gusterson on how the hippies saved physics:
go.nature.com/rzxvhe

were infused with the optimism of the US space programme, responded to talk of scarcity with a visceral aversion. These teens and twenty-somethings latched on to O'Neill's visions of suburbs in space piping abundant solar power and lunar regolith back to Earth. O'Neill himself was ambivalent about their support, and when his star faded they moved on to form or follow other high-tech enthusiast movements, each of which took *The Limits to Growth* as its foil.

McCray's book is especially convincing in following the various movements that arose in reaction to the Club of Rome's 1972 book. At present, we face genuinely alarming limits to growth. Our ability to comprehend and act on such constraints — particularly with respect to climate change and alternative energy — is still distorted by the infelicities in the first edition of *The Limits to Growth* and the ferocious reaction to its conclusions. Some visioneering ideas for overcoming limits to economic growth have contributed to inaction on climate change by promising an appealing but impossibly easy, sacrifice-free, small-government path to a limitless future. These have distracted attention from politically difficult, less technology-intensive solutions.

McCray's argument that visioneers play an important part in the "technological ecosystem" is also compelling, but asymmetrically deployed. For one thing, as the book's subtitle implies, only those who propose a limitless future get to be visioneers; technical experts who popularize visions of a future that is constrained by scarcity (Forrester or the biologist Barry Commoner, for example) apparently do not count. McCray also sometimes treats his visioneers less critically than their foils. He describes *The Limits to Growth* as "refuted" by experts, but treats equally damning arguments against the visions of O'Neill and Drexler in a 'he-said–she-said' fashion. For instance, Nobel Laureate Richard Smalley's contention that Drexler's molecular gears and conveyor belts obey an impossible chemistry is dismissed as "Drexler and Smalley largely talk[ing] past one another".

Yet McCray is correct that visioneers influence, and are influenced by, an ecosystem of philanthropists, politicians, funding agencies, entrepreneurs, undergraduates, scientists and others. That group spurs technological innovation, crafts science policy, and shapes and shares widely held visions of the future. ■

Cyrus C. M. Mody *is an assistant professor in the History Department, Rice University, Houston, Texas.*
e-mail: cyrus.mody@rice.edu

ASTROBIOLOGY

The cosmological you

Birger Schmitz weighs up an exploration of how the Universe permeates us.

Neil Shubin's masterpiece *Your Inner Fish* changed the way I see myself. Using evidence of the first fish-amphibians that left the oceans for land 375 million years ago, Shubin described with stunning clarity how every aspect of our anatomy goes back to our distant ancestors.

Now, in his follow-up *The Universe Within*, he takes the discussion a step further: how the Universe formed, our place in the Solar System and the intertwined evolution of our planet and life. He shows that all this is built into us as physical beings.

Inside us, for instance, are atoms that formed in exploding stars. The movements of heavenly bodies are inherent in our perception of time and in biological clocks. Physical parameters such as gravity determined our shapes and sizes. Had Jupiter formed closer to the Sun, we would have turned out short and squat; farther out and we would have been slender and elongated. This is because Jupiter formed before the inner rocky planets, and its position relative to the Sun determined Earth's size and gravity.

Shubin starts with the formation of the Universe 13.7 billion years ago, segueing into that of the Solar System 4.6 billion years ago. Much later, about 200 million years ago, when the supercontinent Pangaea broke up, the continents and ocean basins we know today began to form. This was accompanied by the rapid evolution of more complex life forms — dinosaurs, mammals and birds.

> ***"The mystery of why we are here is perhaps greater than ever."***

Shubin suggests a rather original connection between continental break-up and the evolution of such creatures: mud settling on the vast stretches of coastline created by the break-up of Pangaea buried biological material that would otherwise have decayed in water, using up oxygen. The result, Shubin says, was an increase in atmospheric oxygen, one of the key factors that allowed animals to conquer land. Mammals require a lot of oxygen to maintain their high-energy, warm-blooded lifestyle. Life on the low-oxygen Earth of 200 million years ago would have been like that today at 4,500 metres above sea level.

Much of the second half of *The Universe Within* summarizes the history of how our geological view of Earth developed. It incorporates stories such as how the discovery of similar fossil organisms on distant continents led Alfred Wegener and others towards the idea of continental drift. We also meet William Smith, who invented stratigraphy, Louis Agassiz, who discovered ice ages, and geologist Bruce Heezen and oceanographic cartographer Marie Tharp, who were central to developing the theory of plate tectonics.

The Universe Within: Discovering the Common History of Rocks, Planets, and People
NEIL SHUBIN
Pantheon: 2013. 240 pp. $25.95, £20

Shubin is at his best when he deals with anatomy and biology, as in his discussion of the inventive geologist Michel Siffre. In 1962, Siffre spent two months living in a subterranean cave to gauge whether he could track time without any tools with which to measure it. After two months, he was convinced that only 37 days had passed. This was in line with what we know about the role in 'internal clocks' of the pineal gland, which regulates the production of sleep-inducing melatonin depending on the available light. Shubin's storytelling in such passages is gripping.

The Universe Within is a charming and enjoyable read, but it does not reach the heights of *Your Inner Fish*. There is a familiar feel to some of the sections, and the book's title raises expectations that are not really met. Where are the mysteries of the brain, the laws of thought and our consciousness? These, to me, are the most amazing aspects of the 'universe within'. In my view, the popular astronomy writer Timothy Ferris has touched on these aspects of the relationship between the soul and the Universe in a more thought-provoking way in books such as *The Mind's Sky* and *The Whole Shebang*.

And what if our view of the Universe continues to change as much as it did in the past century? From Shubin, one gets the impression that much is now solved. But the mystery of why we are here is perhaps greater than ever. Maybe, as the physicist Max Planck put it: "Science cannot solve the ultimate mystery of nature. And that is because, in the last analysis, we ourselves are part of nature and therefore part of the mystery that we are trying to solve." ■

Birger Schmitz *is a professor of geology at Lund University in Sweden and the leader of the ASTROGEOBIOSPHERE project.*
e-mail: birger.schmitz@nuclear.lu.se

Correspondence

NIH funding: agency rebuts critique

We disagree with Joshua Nicholson and John Ioannidis' claim that the peer-review system of the US National Institutes of Health (NIH) works against genuinely innovative research, because we believe that their analysis is flawed (*Nature* **492,** 34–36; 2012).

They use 1,000 or more citations as a proxy for identifying breakthrough discoveries. However, more than 60% of the 158 highly cited articles they analyse cannot be categorized as innovative primary research: 34% are reviews, reports, clinical guidelines or descriptions of resources; 18% are clinical trials of the type primarily funded by industry; and 11% fall outside the NIH mandate that research should have the potential to improve human health (for our analysis, see http://dpcpsi.nih.gov/opa/natcorr).

Excluding these articles leaves 58 of the original 158: of these, 83% were funded by the NIH and 17% were funded by private industry.

The NIH welcomes further innovation in biomedical research, as evidenced by funding mechanisms such as the NIH Director's Pioneer, Transformative Research, Early Independence and New Innovator awards.

George Santangelo *Office of Portfolio Analysis, National Institutes of Health, Bethesda, Maryland, USA.*
george.santangelo@nih.gov
David J. Lipman *National Center for Biotechnology Information, National Institutes of Health, Bethesda, Maryland, USA.*

NIH funding: it does support innovators

Joshua Nicholson and John Ioannidis conclude that "too many US authors of the most innovative and influential papers in the life sciences do not receive NIH funding", on the basis of their analysis of 200 papers sampled from 700 life-sciences papers with 1,000 or more citations (*Nature* **492,** 34–36; 2012). However, my reanalysis of their data suggests that the US National Institutes of Health (NIH) has supported a substantial proportion of such contributions over the past 12 years.

For a random sample of 125 of the authors' pool of 700 highly cited papers, I found that 63 were reviews and 17 fell outside the life sciences. Of the remaining 45 original research papers, 34 (that is, 75%) were supported by the NIH (source: NIH grants database and my own analysis); the other 11 papers did not receive NIH support for various reasons (for details, see go.nature.com/nywiid).

Nicholson and Ioannidis further underestimate the NIH's support for groundbreaking research by requiring both the first and last authors to be principal investigators on an NIH grant when, in fact, first authors are often graduate students.

Steven L. Salzberg *Johns Hopkins School of Medicine, Baltimore, Maryland, USA.*
salzberg@jhu.edu

NIH funding: the critics respond

In re-analysing our data (*Nature* **492,** 34–36; 2012), George Santangelo with David Lipman, and Steven Salzberg each exclude two-thirds of the top-cited US publications assigned by Scopus to life or health sciences. We consider that their analyses discard most health research that matters.

Clinical trials are primary research that catalyses preventive or therapeutic innovation; reports and guidelines also decisively inform and radically transform health. Extremely highly cited reviews formulate pioneering concepts or synthesize influential work. Also, scholars reaching the top 0.01% with papers of any type predict some further excellence.

The authors' analyses depend on grant acknowledgments (75–83% in both series) but these are a problematic metric because they represent grants that were awarded 10–15 years ago when the NIH budget was expanding and acceptance rates were highest. Also, any of several co-authors can acknowledge a funding source if they are under pressure to demonstrate grant-related productivity, even if that source is irrelevant. And let's imagine a hypothetical funding system that forces all geniuses to quit science: 100% of papers could still acknowledge funding.

Groundbreaking projects account for less than 1% of awarded grants. Students whose papers reach 1,000 or more citations should certainly become principal investigators: stars will abandon systems that stifle independence.

John P. A. Ioannidis *Stanford Prevention Research Center, Stanford, California, USA.*
jioannid@stanford.edu
Joshua M. Nicholson *Virginia Tech, Blacksburg, Virginia, USA.*

Science alone cannot shape sustainability

We agree that science can and should inform the sustainable development goals agreed at the Rio+20 conference (G. Glaser *Nature* **491,** 35; 2012). But basing policy decisions on science alone may be unproductive.

Experience indicates that political receptiveness to scientific advice is essential for shaping policy. Climate-change policy-makers, for example, continually receive new technical and scientific input, but they still cannot agree how best to act on it to mitigate or adapt to climate change (J. Depledge *Glob. Environ. Polit.* **6,** 1–22; 2006).

Scientific evidence is rarely politically neutral or universally accepted; neither can it replace what is inherently a political process.

Suraje Dessai, Stavros Afionis, James Van Alstine *University of Leeds, UK.*
s.dessai@leeds.ac.uk

Prize marks German journal centenary

One hundred years ago today, the publisher Springer launched a German multidisciplinary journal named *Naturwissenschaften* ('natural sciences') at the behest of physicist and future editor Arnold Berliner. The journal was closely modelled on *Nature*.

Nature maintained its support for the journal throughout the political upheavals in twentieth-century Europe. As a Jew, Berliner was forced to resign in 1935. *Nature* wrote: "We much regret to learn that on August 13 Dr. Arnold Berliner was removed from the editorship of *Die Naturwissenschaften*, obviously in consequence of non-Aryan policy" (*Nature* **136,** 506; 1935). It also published a moving obituary when Berliner took his own life in 1942 (*Nature* **150,** 284; 1942).

Naturwissenschaften — The Science of Nature is now published exclusively in English. During this anniversary year, the journal will provide open access online to its most highly rated 100 articles, and will present the first Arnold Berliner Award for the best research article published in 2012, judged using Berliner's original motivations of excellence, originality and interdisciplinarity.

Joan Robinson *Springer, Heidelberg, Germany.*
joan.robinson@springer.com

CONTRIBUTIONS

Correspondence may be sent to **correspondence@nature.com** after consulting the guidelines at **http://go.nature.com/cmchno**.

FORUM Palaeontology

Fossils come in to land

Fossils found in rocks of the Ediacaran period in Australia have been previously characterized as early marine organisms. But a report suggests that these rocks are fossilized soils. So did some of these Ediacaran organisms in fact live on land, like lichens? A palaeontologist and a geologist weigh up the evidence. SEE LETTER P.89

THE PAPER IN BRIEF

- The Ediacaran period, 635 million to 542 million years ago, immediately predates the Cambrian period, which saw an evolutionary explosion that led to all modern animal phyla.
- Fossils from an Ediacaran geological formation in South Australia have been classified as invertebrates, protists or fungi, but they have invariably been thought of as being marine.
- Retallack proposes, in a paper published in this issue (page 89)[1], that the Ediacara Member contains fossilized soils (palaeosols)*.
- The presence of palaeosols suggests that some of the fossils within them may have been lichen-like organisms or microbial colonies that lived on land, rather than in the ocean.

Muddying the waters

SHUHAI XIAO

Fossils in the Ediacara Member in South Australia have been traditionally interpreted as representatives of ancestral marine organisms[2]. But, breaking away from this tradition and pursuing his own radical interpretation[3,4], Retallack[1] now proposes that these fossiliferous beds are palaeosols and that some Ediacaran fossils are soil lichens or colonies of soil microbes. These propositions would represent a fundamental change in our picture of evolution, but they will probably face continuing scepticism because the evidence is unconvincing.

Definitive identification of palaeosols in the Ediacara Member is a challenge, because this unit was deposited before land plants arose and it thus lacks features that are diagnostic of ancient soils, such as traces of plant roots. But Retallack cites a host of observations as evidence that the fossiliferous Ediacara Member originated from soil formation (pedogenesis): its reddish colour, its elemental and stable-isotopic geochemistry, patterns of surface disruption, and the presence of sand crystals of gypsum and nodules of carbonate.

However, the evidence is ambiguous. For example, the reddish colour and depletion of certain elements in the Ediacara Member could be a result of weathering that occurred during the Cenozoic era (from 65 million years ago to the present), rather than resulting from chemical weathering of the rocks through pedogenesis during the Ediacaran period[5]. Retallack counters that Cenozoic weathering would have produced continuously reddish strata, rather than the observed alternation of grey and red beds, but he fails to recognize that weathering colours can vary with lithological characteristics of the rocks (such as mineralogical composition and permeability). In addition, carbonate nodules and sand crystals of gypsum are common features of marine sediments, and the isotope signatures of carbonate nodules in the Ediacara Member can be accounted for by post-depositional alterations that do not involve pedogenic processes.

> ***"This ambiguous evidence for pedogenesis is outweighed by compelling evidence for the marine origin of the Ediacara Member."***

Retallack further illustrates his argument for palaeosols with examples of large-scale disruption structures characteristic of soils (see Fig. 2b of the paper[1]), but such structures are intriguingly similar to slumps or load structures resulting from subaqueous and post-depositional movement of sediments. He also depicts small-scale disruption structures, which he interprets as having been caused by millimetre-sized tubules that might be the fossilized remains of bacterial filaments, lichen rhizines (root-like filaments) or fungal hyphae (see Fig. 2c–g of the paper[1]). But I find this interpretation dubious, because the tubules are too irregular to be confidently interpreted as being derived from microbes.

In my opinion, this ambiguous evidence for pedogenesis is outweighed by compelling evidence for the marine (or at least subaqueous) origin of the Ediacara Member. For example, benthic Ediacaran organisms (those that lived on or within sediments), such as *Cyclomedusa davidi* and *Dickinsonia costata,* are preserved *in situ* on rippled bedding surfaces[5] (Fig. 1). In addition, some Ediacaran fossils show holdfasts (root-like structures) that were dragged in the same direction as the alignment of attached stalks[6]. These features could not have formed without the action of waves or currents. And detailed sedimentological analysis has revealed a suite of features characteristic of subaqueous deposition[5], including ripple marks and current lineations.

A palaeosol interpretation leads Retallack to reinterpret fossils in the Ediacara Member as the remains of soil lichens, microbial colonies, fungi, slime-mould trails or casts of needle ice (which forms in frozen soil). However, many Ediacaran species in Australia are also found worldwide in unambiguously marine formations, such as black shales and limestones. Furthermore, the Ediacaran fossil *Dickinsonia* shows evidence of intermittent locomotion — but lichen do not move. The Ediacaran fossil *Radulichnus*, interpreted as casts of needle ice by Retallack[1] but as traces of grazing organisms by others, has fanning sets of parallel scratches, an arrangement that cannot be explained by needle ice. And although Retallack proposes that Ediacaran trace fossils are trails left by land-cruising slugs or aggregating slime moulds, these organisms could not have made the burrows that are clearly visible in the Ediacara Member.

On a positive note, Retallack's persistent pursuit of the idea of soils and lichens does motivate us to rethink the possibility that lichens, whether terrestrial or marine[7], might have existed at this early time, and played a part in regulating the Ediacaran Earth before the rise of vascular plants. But we need

*This article and the paper under discussion[1] were published online on 12 December 2012.

J. GEHLING/SOUTH AUSTRALIAN MUSEUM

Figure 1 | Sand or soil. The picture shows Ediacaran fossils (*Dickinsonia costata*) on a rippled surface, found in the Ediacara Member in South Australia. Previous interpretations suggest that these fossils represent marine organisms that lived on the sea floor[5], but Retallack[1] proposes that they were land dwellers.

clearer evidence before we should consider redrawing the timeline of life's transition from sea to land.

Shuhai Xiao *is in the Department of Geosciences, Virginia Tech, Blacksburg, Virginia 24061, USA.*
e-mail: xiao@vt.edu

Not all at sea

L. PAUL KNAUTH

The last 700 million years of the Precambrian eon was apparently a time of major sea-level changes in which continental margins were alternately inundated and then left exposed to land-surface erosion. Ancient soils were being eroded into the sea[8], and some should have been preserved as palaeosols similar to those in strata from more recent times. But so far, reports of late Precambrian examples are lacking. This may simply reflect the difficulty of recognizing soil strata without the root traces so obvious in much younger rocks. Other diagnostic features are more subtle and are generally recognized only by those experienced in studying palaeosols. Retallack is one such specialist, and he is well positioned to argue for the existence of soil-inhabiting Ediacaran organisms.

Retallack has long pondered the nature of Ediacaran organisms, and controversially proposed that they were actually not animals but large lichens[3]. This heterodox interpretation would be greatly strengthened if he really has found examples that lived on land. Whether animal or lichen, the discovery would indicate that some organisms mastered the transition from marine to non-marine life much earlier than currently thought — or even support the possibility that the transition went the other way, to ultimately account for the Cambrian explosion in the sea[9].

So how strong are his arguments? Interpreting ancient depositional environments is a tricky business, and a stratigraphic layer without telltale root fossils may be a palaeosol only in the eye of the beholder. For example, what Retallack suggests are geological relics of soil deformation have been suggested by others to be water-escape features[10]. His cogent arguments that the red colour represents Precambrian weathering will be resisted by those familiar with the extensive red colour imparted to Australian rocks during modern weathering. In the absence of comparisons with Ediacaran examples that Retallack accepts as marine, it is debatable whether a lack of non-overlapping specimens really precludes wash-ups on a shore, as he suggests. Furthermore, the isotope data for the carbonate nodules that Retallack claims represent subaerial exposure are also compatible with coastal recharging of rainwater into subsurface aquifers known commonly to extend far offshore[11]. And finally, it is difficult to distinguish the sedimentary structures that the author interprets as sand deposited in a terrestrial valley from what could belong to a submarine canyon, as has been proposed[5].

"His considered case means that researchers sceptical of his interpretations will need to become experts in palaeosol characteristics."

As is usual in sedimentology, observations can be construed in alternative ways, and interpretations for these strata have historically covered the gamut of geological possibilities — from lacustrine to lagoonal, coastal and open marine. It is appropriate that interpretations change or are superseded with the arrival of new observations, and that is why this publication is fascinating and timely and should be considered seriously. Although Retallack's ideas are at odds with the accepted dogma, these do not need to be mutually exclusive. Why should it not be possible that some Ediacaran organisms lived on land, even if most of the other sites in which they have been found are interpreted as marine? There is still uncertainty regarding exactly what kind of organisms they were, so eliminating a possible habitat on the basis of whether or not they are animals is unwarranted. Retallack considers new data and observations and provides comprehensive reasoning for each of his points. His considered case means that researchers sceptical of his interpretations will need to become experts in palaeosol characteristics to mount convincing counter-arguments.

The search for late Precambrian biological evolution in the non-marine realm is an exciting new frontier, especially considering carbon-isotope data that probably indicate a late Precambrian greening of land surfaces[12]. Ediacaran organisms living in soils would be further evidence that land areas in this interval of Earth's early history were not biologically barren surfaces as is commonly assumed. We were not there when all this happened and will never know for certain what actually happened when. So I say, until the forensic evidence for Ediacaran habitats becomes strongly compelling one way or the other, let multiple hypotheses thrive! ■

L. Paul Knauth *is in the School of Earth and Space Exploration, Arizona State University, Tempe, 85287-1404 Arizona, USA.*
e-mail: knauth@asu.edu

1. Retallack, G. J. *Nature* **493,** 89–92 (2013).
2. Xiao, S. & Laflamme, M. *Trends Ecol. Evol.* **24,** 31–40 (2009).
3. Retallack, G. J. *Paleobiology* **20,** 523–544 (1994).
4. Retallack, G. J. *Sedimentology* **59,** 1208–1236 (2012).
5. Gehling, J. G. *Precambr. Res.* **100,** 65–95 (2000).
6. Tarhan, L. G., Droser, M. L. & Gehling, J. G. *Palaios* **25,** 823–830 (2010).
7. Yuan, X., Xiao, S. & Taylor, T. N. *Science* **308,** 1017–1020 (2005).
8. Kennedy, M., Droser, M., Mayer, L. M., Pevear, D. & Mrofka, D. *Science* **311,** 1446–1449 (2006).
9. Knauth, L. P. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **219,** 53–69 (2005).
10. Callow, R. H. T., Brasier, M. D. & McIlroy, D. *Sedimentology* http://dx.doi.org/10.1111/j.1365-3091.2012.01363.x (2012).
11. Hathaway, J. C. *et al. Science* **206,** 515–527 (1979).
12. Knauth, L. P. & Kennedy, M. J. *Nature* **460,** 728–732 (2009).

COMPUTATIONAL MATERIALS SCIENCE

Soft heaps and clumpy crystals

A detailed simulation of the packing behaviour of deformable particles settles the debate about whether soft matter can adopt an unconventional crystal structure at high densities — it can. The hunt is now on for a real-world example.

FRANCESCO SCIORTINO & EMANUELA ZACCARELLI

Imagine the Rome metro at rush hour: passengers are squeezed into close contact with one another. But there is a physical limit beyond which they cannot go, because their bodies cannot occupy the same space. This common experience has an equivalent at the molecular scale, in what physicists call excluded volume — strong repulsive forces, of quantum-mechanical origin, that prevent atoms from occupying the same space. Because of this phenomenon, dense arrangements of atoms and molecules result in solids that have lattice structures, in which each particle excludes neighbours from its site in the lattice. It is therefore surprising to read Lenz and colleagues' paper[1] in *Physical Review Letters*, which reports that the packing of soft particles may result in unusual crystals in which each lattice site is occupied not by a single particle, but by clumps of particles.

Soft particles are nanometre- or micrometre-sized macromolecules that have a deformable shape. Focusing on polymers, for example, one can envisage several different soft particles of increasing complexity (Fig. 1). These could be: linear chains; rings, in which the ends of a polymer chain are connected; stars, in which several polymer chains are joined at a common centre; and dendrimers, in which several stars are linked together. For a fairly small energy cost, the structures of soft particles can rearrange and interpenetrate to cope with excluded-volume constraints at the molecular scale. This allows the centres of mass of different soft particles to coincide, without any overlapping of the monomers (Fig. 1).

To predict the collective behaviour of soft particles, scientists typically use one of two theoretical approaches. The first approach is to perform calculation-intensive, monomer-resolved simulations that provide an accurate description of a system's properties but are at the limit of today's computational capabilities. The second is to use the tools of statistical mechanics to develop a simplified (coarse-grained) model of a soft particle and its interactions with other particles. Such simplifications are often so crude that each particle is represented as a single site at the particle's centre of mass, but they allow predictions to be derived more easily.

Theoretical investigations using coarse-grained models suggest that, in the dense fluid state[2], soft particles prefer to sit on top of each other, forming heaps or clumps. Indeed, beyond a certain density it becomes energetically preferable for a particle to completely overlap with a few others, rather than be subjected to the cumulative repulsion of many more neighbours (Fig. 2). Similarly, passengers on the metro could consider piling up on top of one another rather than suffering from excessive squeezing. In fact, parents often do this with their children, by holding them in their arms!

Further work[3] has shown that if such a 'clumpy' fluid crystallizes, the resulting solid retains the clumpy structure to form a regular arrangement of heaps. The number of particles in each heap varies, giving rise to solids that are spatially ordered (as is typical of standard crystals), but also locally disordered because of the random number of particles in the clumps. However, predictions based on coarse-grained models are built on approximations that typically become progressively less accurate as the particle system gets denser, and so are open to question. The prediction of clumpy crystals from coarse-grained models has therefore been seen by some as an academic curiosity, especially given that monomer-resolved simulations of polymer rings[4] did not confirm the predicted formation of such crystals.

Lenz and colleagues' monomer-resolved simulations for dendrimers now provide unmistakable evidence that clumpy crystals really can form. The authors' work builds on lessons learned from the earlier computational study of polymer rings[4], which revealed that the particles shrank as density increased. The shrinkage progressively invalidated predictions made using coarse-grained simulations. To overcome this problem, the authors modelled dendrimers that have a dense core, which prevents the particles' size from varying significantly as particle packing increases. The researchers' monomer-resolved simulations of the dendrimers closely follow the theoretical predictions from corresponding coarse-grained models — that is, they confirm that clumpy crystals can develop.

The time is now ripe for an experimental search for clumpy crystals composed of specially synthesized soft particles. DNA dendrimers[5] — nanometre-scale particles designed to self-assemble from single-stranded DNA molecules — may be the optimal candidates for realizing this unconventional state of matter. If so, the resulting clumpy crystals would enter the fast-growing pantheon of DNA constructs with potential uses in nanotechnology[6].

Clumpy crystals shed light on fundamental physical principles, but they may also have practical applications. The variation in the number of particles at each lattice site favours mass transport, in the form of individual particles hopping from one site to another (Fig. 2). This could be valuable for applications in which lattice rigidity and mass transport need to be coupled. Another fascinating possibility is that the number of particles occupying clumps in a crystal could be controlled by compressing the crystal. Indeed, a theoretical study has suggested[7] that a sequence of crystal phases, with differing occupancy numbers, occurs as the density of the particles increases.

The peculiar nature of clumpy crystals should be shared by disordered solids — those that do not have lattice structures — such as glasses. Glassy states[8] offer unique opportunities to tailor the viscoelastic properties of

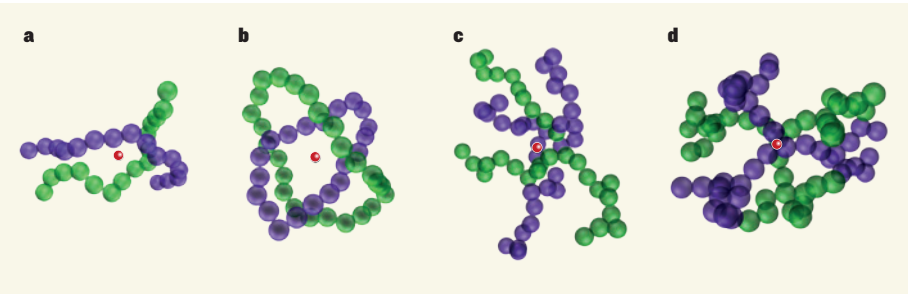


Figure 1 | Two particles, one centre. **a–d**, Soft particles formed from polymers can change shape and interpenetrate in such a way that their centres of mass coincide, as shown here for pairs of: polymer chains (**a**); polymer rings (**b**); three-armed polymer 'stars' (**c**); and dendrimers (connected stars; **d**). The centres of mass (red dots) of the green and purple particles coincide, even though not all of the monomers overlap. (Graphic courtesy of Lorenzo Rovigatti.)

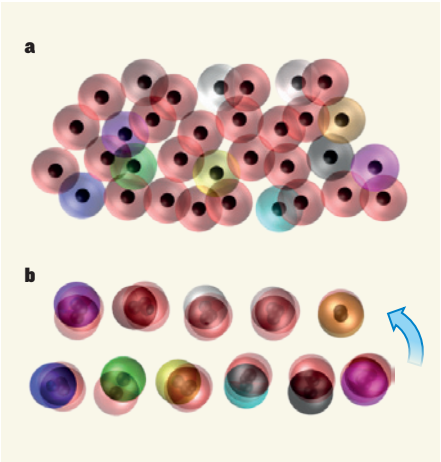


Figure 2 | The formation of clumpy crystals in two dimensions. **a**, At high density, soft particles (shown as transparent spheres; diameter corresponds to each particle's typical size) might adopt an arrangement in which they partially overlap with several neighbours. The total repulsion exerted on each particle by its neighbours is high. **b**, Alternatively, the same particles might form a regular lattice of 'clumps'; in this case, each clump contains an average of three overlapping particles. Particles belonging to distinct clumps do not interact, so that the overall repulsion exerted on each particle is less than that in **a**. Particles might also be able to hop between lattice sites (arrow). Lenz and colleagues' numerical simulations[1] reveal that dendrimeric soft particles form clumpy crystals. (Graphic courtesy of Lorenzo Rovigatti.)

materials. For example, glasses can be melted by applying a 'shear' deformation force parallel to a sample's surface (an effect known as shear melting). Analogous to what has been reported for clumpy crystals[9], the viscosity of shear-melted clumpy glasses should increase with the intensity of the applied deformation, a phenomenon known as shear thickening. This behaviour is at odds with that of most materials, in which the application of shear decreases a sample's viscosity. Finally, the link between soft particles and quantum mechanics should be noted: boson particles have been predicted[10] to form supersolids, the quantum analogue of clumpy crystals, although such supersolids have not yet been observed. Perhaps an experimentally realized clumpy crystal could give insight into some aspects of such mysterious quantum solids.

The unconventional behaviour of soft matter has often surprised scientists. Lenz and colleagues' study provides yet another example of how soft particles at the nano- and microscale do not simply reproduce phenomena known to occur in the atomic and molecular world. ■

Francesco Sciortino *is in the Department of Physics, Sapienza Università di Roma, Rome I-00185, Italy.* **Emanuela Zaccarelli** *is at CNR-ISC, Institute of Complex Systems, Rome I-00185, Italy.*

e-mails: francesco.sciortino@uniroma1.it; emanuela.zaccarelli@cnr.it

1. Lenz, D. A., Blaak, R., Likos, C. N. & Mladek, B. M. *Phys. Rev. Lett.* **109,** 228301 (2012).
2. Klein, W., Gould, H., Ramos, R. A., Clejan, I. & Mel'cuk, A. I. *Physica A* **205,** 738–746 (1994).
3. Mladek, B. M., Gottwald, D., Kahl, G., Neumann, M. & Likos, C. N. *Phys. Rev. Lett.* **96,** 045701 (2006).
4. Narros, A., Moreno, A. J. & Likos, C. N. *Soft Matter* **6,** 2435–2441 (2010).
5. Li, Y. *et al. Nature Mater.* **3,** 38–42 (2004).
6. Seeman, N. C. *Nature* **421,** 427–431 (2003).
7. Zhang, K., Charbonneau, P. & Mladek, B. M. *Phys. Rev. Lett.* **105,** 245701 (2010).
8. Coslovich, D., Bernabei, M. & Moreno, A. J. *J. Chem. Phys.* **137,** 184904 (2012).
9. Nikoubashman, A., Kahl, G. & Likos, C. N. *Phys. Rev. Lett.* **107,** 068302 (2011).
10. Cinti, F. *et al. Phys. Rev. Lett.* **105,** 135301 (2010).

ASTRONOMY

Andromeda's extended disk of dwarfs

Deep-imaging observations of the Andromeda galaxy and its surroundings have revealed a wide but thin planar structure of satellite galaxies that all orbit their host in the same rotational direction. SEE LETTER P.62

R. BRENT TULLY

In this issue, Ibata and colleagues[1] report that roughly half the dwarf companion galaxies of the Andromeda galaxy are rotating coherently about it in a thin plane. Their finding provides a fascinating new constraint on theories of galaxy formation.

First, the observational facts. Andromeda, also known as Messier 31, is the nearest giant galaxy to our Galaxy. It is so near that a census of its companions by deep imaging with the Canada–France–Hawaii Telescope has been completed over a large area and to a faint level of detection. Because the system is close, distances to the companions can be measured and the velocities of constituent stars determined. Such a complete census for the Milky Way is impossible because candidates could be anywhere in the sky, even behind a zone obscured by the Galaxy. And other giant galaxies are too far away to be studied to such a level of detail.

Ibata *et al.* found that 13 of 27 dwarf companions (satellites), at distances from Messier 31 of between 35 and 400 kiloparsecs (114–1,305 light years), lie in a thin plane 13 kpc thick and share a coherent velocity pattern: those to the north of Messier 31 are moving away from Earth relative to the galaxy, and those to the south are moving relatively towards Earth. No theorist of galaxy formation would have dared to predict such a situation. What's more, the Milky Way is in the same plane as the 13 satellites. The discovery of this plane is a spectacular result, and the authors avoid the risk of diluting their message by not mentioning more speculative matters that add to the intrigue.

Although the disk of Messier 31 is tilted by about 50° from the plane of the satellites, the rotation in the galaxy is in the same direction of motion as the satellite-velocity pattern. Looking beyond Ibata and colleagues' survey region, the three galaxies that lie 250–500 kpc from Messier 31 — IC 1613, IC 10 and LGS 3 — and which were known before the advent of deep-imaging surveys, all reside in the same satellite plane. This is particularly interesting because these three more-distant galaxies contain substantial interstellar gas and are still forming stars, and so might be recent arrivals on the scene. All the satellites in the authors' survey region are gas deficient (except Messier 33, which is not on the plane being discussed) and, according to standard galaxy-formation models, would be presumed to have been in the vicinity of Messier 31 for some time and to have complex orbits.

But things are even stranger than Ibata and colleagues suggest. The remaining known Messier 31 satellites can be split roughly in equal numbers into those at lower and higher Galactic longitude. All of those at higher longitude than Messier 31, including the Local Group's third-largest galaxy Messier 33, lie in a separate common plane. This secondary plane is offset and tilted by about 13° from the primary plane through Messier 31.

Ibata *et al.* remark on earlier suggestions that satellites of the Milky Way also seem to lie in a plane[2,3]. It occurred to me that perhaps there was enough information in the data archives to evaluate the distribution of companions in the next-nearest groups of galaxies — those around the dominant galaxies Centaurus A and Messier 81. The regions around these galaxies have been closely studied in surveys for satellite candidates and in follow-up observations with the Hubble Space Telescope[4,5]. In the case of Centaurus A, 22 of 24 satellites within 600 kpc of the galaxy's centre separate into two equally populated planes that are roughly parallel but offset by 280 kpc. Centaurus A lies in

one of these planes. Most of the companions to Centaurus A are gas poor, but there are several systems that contain gas and in which star formation is occurring in each plane.

The situation in the Messier 81 Group is less compelling but still suggestive. Here, there is a distinction between the distribution of the gas-poor satellites and that of gas-rich satellites that are undergoing star formation. The gas-poor systems lie in a flattened distribution that have characteristic dimensions of 60 × 120 kpc, with the flattening coincident with the 'Local Sheet' structure[6] that harbours all the galaxies mentioned above and which extends over a long dimension of 10 Mpc and with a thickness of 1 Mpc. The gas-rich satellites typically lie farther from Messier 81, and loosely align to a plane of their own.

This discussion of the organized distribution of satellites is anchored in the solid evidence reported by Ibata and colleagues for a thin plane with coherent kinematics. There are hints that structure in the distribution of satellites is the norm. The subject deserves further attention, but it should be noted that the planes that have been discussed on scales of 300–500 kpc have a general alignment with the Local Sheet. This sheet forms a wall of an anti-structure, the 'Local Void', that strongly affects the development of nearby structure[6].

Ibata *et al.* only touch on possible scenarios underlying the formation of the planar structures. The new information compounds a familiar galaxy-formation problem — a deficiency in the numbers of satellites found compared with theoretical expectations[7,8]. Now, it seems, not only is there a paucity of satellites, but also most of those that do exist are in these organized structures. The very organization suggests that the structures (possibly as distinguished from their constituents) are not ancient.

Current ideas about galaxy formation propose that material (both gas and already constructed galaxies) falls into the extended haloes around galaxies as flows along filaments. The orbital angular momentum of the infalling material over time tends to cause motion that has the same direction of rotation as that of the dominant galaxy in the halo, resulting in the build-up of a spiral disk in the galaxy. It is reasonable to assume that newly accreted satellites would share the sense of rotation, but that after a few orbits they would tend to become scrambled. Because infalling galaxies around Messier 31 adhere to such a thin plane, it would seem that they do not take many excursions before they are absorbed in the central galaxy. ■

R. Brent Tully *is at the Institute for Astronomy, University of Hawaii, Honolulu, Hawaii 96822, USA.*
e-mail: tully@ifa.hawaii.edu

1. Ibata, R. A. *et al. Nature* **493,** 62–65 (2013).
2. Lynden-Bell, D. *Mon. Not. R. Astron. Soc.* **174,** 695–710 (1976).
3. Pawlowski, M. S., Pflamm-Altenburg, J. & Kroupa, P. *Mon. Not. R. Astron. Soc.* **423,** 1109–1126 (2012).
4. Karachentsev, I. D. *et al. Astron. Astrophys.* **385,** 21–31 (2002).
5. Chiboucas, K., Karachentsev, I. D. & Tully, R. B. *Astron. J.* **137,** 3009–3037 (2009).
6. Tully, R. B. *et al. Astrophys. J.* **676,** 184–205 (2008).
7. Klypin, A., Kravtsov, A. V., Valenzuela, O. & Prada, F. *Astrophys. J.* **522,** 82–92 (1999).
8. Moore, B. *et al. Astrophys. J.* **524,** L19–L22 (1999).

DEVELOPMENTAL BIOLOGY

Segmentation within scale

Irrespective of an organism's size, the proportional sizes of its parts remain constant. An experimental model reveals size-dependent adjustment of segment formation and gene-expression oscillations in vertebrates. SEE LETTER P.101

NAAMA BARKAI & BEN-ZION SHILO

Developing organisms face a major challenge: their body pattern must be adjusted — scaled — to their body size. But how is tissue size 'measured'? And what conveys general size information to the local settings of each cell? Despite intense interest, the mechanistic basis of scaling is poorly understood. On page 101 of this issue, Lauschke *et al.*[1] report that scaling persists in a tissue-culture model that simulates early segmentation in the vertebrate embryo. The simple, two-dimensional geometry of this system, and the fact that it can be visualized in real time and manipulated, opens exciting avenues for studying the formation and scaling of vertebrate segmentation*.

The segmented organization of vertebrates is set up in the early embryo. As the embryo elongates along an anterior–posterior axis, segmented structures called somites bud regularly from the anterior end of its immature presomitic mesoderm (PSM) tissue[2–4]. The number of segments differs between species, but varies little between individuals of the same species. Seminal work by the developmental biologist Jonathan Cooke showed that surgical manipulations that reduce embryo size generate smaller yet well-proportioned embryos that are patterned normally along both anterior–posterior[5] and dorso-ventral[6] axes. In particular, somites become proportionally smaller, but their number and relative position are maintained[5].

The observation that somite number and size are regulated independently prompted the 'clock and wavefront' model, which postulates that spatial and temporal inputs are combined to define somite size and position[7,8]. According to this model, the position at which a somite can be formed at a given time is defined by molecular concentration gradients that are positioned at a fixed distance from the posterior pole (the wavefront), and that move posteriorly through the PSM towards the pole as the embryo elongates. In parallel with this, cell-autonomous oscillations in gene expression that are coordinated across the tissue define the timing of segment formation. The overall outcome is sequential segment formation in an anterior-to-posterior direction.

The predicted oscillations in gene activity have been visualized in chemically fixed and in live embryos, and correlate with the progressive pattern of somite formation. Furthermore, genetic manipulations of wavefront velocity or oscillation frequency modulate segment size[3], as predicted by the clock and wavefront model.

In the intact embryo, oscillations are synchronized between adjacent cells, probably through the activity of the Notch signalling pathway. The frequency of oscillation in gene expression decreases towards the anterior PSM, so that anterior cells reach maximal signalling activity later than posterior cells. Therefore, the pattern of Notch activity seems to propagate from the posterior to the anterior PSM. This wave of molecular activity is not part of the original clock and wavefront model. So, what could be the function of such waves? Do they contribute to somite differentiation or scaling? And what is the molecular basis for these dynamics? Answering these and related questions is greatly facilitated by the ability to visualize[9] and perhaps perturb the differentiation process as it progresses[10].

Although methods for live imaging of intact embryos have been developed, they are limited, in part because of the complex geometry of the embryo. Lauschke *et al.* present an *ex vivo* (tissue culture) model that recapitulates

*This article and the paper under discussion[1] were published online on 19 December 2012.

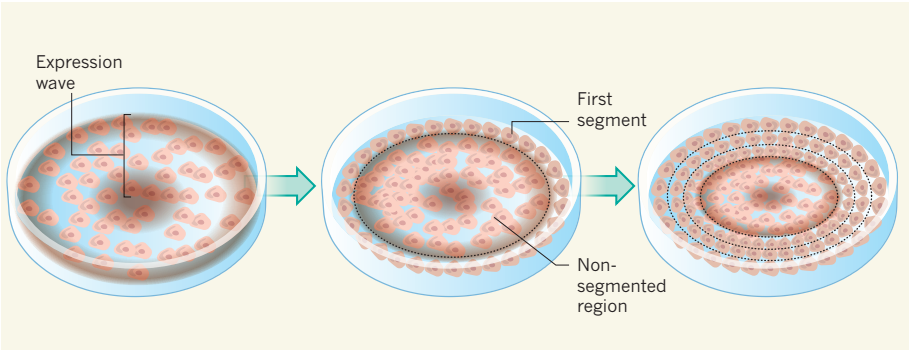


Figure 1 | Oscillating waves of gene expression in culture. Lauschke *et al.*[1] obtained presomitic mesoderm (PSM) tissue from the posterior region of a mouse at day 10.5 of embryonic development. They then plated this sample, which carried a Notch-signalling reporter molecule, on a culture dish. The reporter reveals waves of gene expression that progress in a central-to-peripheral direction. Some 20 hours after plating the cells, segments begin to form from the periphery towards the centre. Notably, segment size is adjusted (scaled) with the decreasing number of available cells, so that each new segment encompasses 20% of the remaining non-segmented region. Gene expression in cells at the centre and periphery oscillates at the same phase (darker shade of brown). Thus, the phase gradient between adjacent cells becomes steeper as the tissue size decreases.

the segmentation process in a simple two-dimensional geometry. They took a tissue slice from the posterior PSM of a mouse embryo and maintained it in culture. Cells grew out from this cultured tissue as a monolayer and began to show the hallmarks of segmentation. Most notably, periodic waves of a 'reporter' molecule for Notch activity seemed to propagate from the centre of the monolayer towards the periphery. After the tissue had grown to a certain size, segments began to form from the periphery, in a temporal sequence that was coordinated with the wave-like activity of the Notch reporter (Fig. 1). Altogether, the authors could detect up to 15 oscillations, at approximately 140-minute cycles, and the formation of five or more segments.

Strikingly, segment scaling was maintained under these *ex vivo* conditions. Formation of a segment decreased the size of the remaining unpatterned tissue. Consequently, the subsequent segments that formed within this smaller region had correspondingly smaller sizes — fixed at 20% of the remaining non-differentiated cells. Scaling was also observed in the velocity of the oscillatory waves of gene expression, which decreased in proportion to unpatterned-tissue size; this meant that the time it took the waves to propagate from the centre to the differentiation front in the periphery remained constant throughout the process.

The authors' further analysis revealed that the best predictor of segment size is the phase gradient, namely, the rate at which the oscillation phase changes across the differentiating tissue. But what determines the phase gradient, and how does it scale with tissue size? The phase gradient is a kinetic property that is not directly linked to any physical entity. As such, scaling based on the phase gradient is different from scaling based on molecular gradients studied previously[11–14]. The correlation between phase gradient and segment size may imply a causal relationship. Alternatively, it could result from a mutual dependence on some other factor.

Perhaps in support of the latter possibility, Lauschke *et al.* observed a temporal increase in the steepness of the phase gradient even before segments had begun to form. One possibility is that both the phase gradient and segment size are dictated by gradients of Fgf or Wnt — signalling molecules that regulate development. The authors report that gradients of these morphogens are indeed established across the tissue, but further quantification is required to determine whether the gradients scale with tissue size, and to analyse their potential role in defining the phase gradient and segment formation.

Whether the segment scaling that Lauschke and colleagues observe in their *ex vivo* system reflects the scaling mechanism that compensates for size variations in live embryos deserves further investigation. Nevertheless, the intriguing *ex vivo* dynamics impinges on fundamental aspects of the clock and wavefront model. How are the clock genes of individual cells adjusted to generate a culture that shows coordinated oscillations in gene expression? Does this coordination involve long- or short-range interactions between the cells? What determines the initiation of segment formation? What is the role of Fgf and Wnt in this process? The authors' model system holds the promise of providing answers to these questions, as well as additional quantitative insight into the process of segment formation in the vertebrate embryo. ■

Naama Barkai *and* **Ben-Zion Shilo** *are in the Department of Molecular Genetics, Weizmann Institute of Science, Rehovot 76100, Israel.*
e-mail: naama.barkai@weizmann.ac.il

50 Years Ago

'Education and the humanist revolution'. By Sir Julian Huxley — The knowledge explosion of the past hundred years has given us a new vision of human identity — of the world, of man, and of man's role in the world … It leads inevitably to a new dominant organization of thought and belief, and, after centuries of ideological fragmentation, to a new comprehensive idea-system, which I call 'evolutionary humanism' … Our new system must itself be evolutionary, not change-resistant but change-promoting; it must transform as well as transmit. In part, it can be achieved by making girls and boys understand the moral duty of helping and guiding the evolutionary process in a desirable direction. But something more practical is also needed. If our aim be greater fulfilment, the next step in psychosocial evolution must be from the Welfare State towards a 'Fulfilment Society'. A humanist educational system will put the idea of a fulfilment society before children, and will provide them with opportunities for actual personal fulfilment.
From *Nature* 5 January 1963

100 Years Ago

Perfect Health For Women and Children. By Elizabeth S. Chesser — Miss Chesser has to be commended for having treated a wide subject in such a sound, common-sense and practical manner as will make the book appeal to every class of reader, both lay and medical. The author does not mince matters when she finds fault with the unhygienic practices of the present day; and the work is full of good, telling sentences, such as, "if women paid as much attention to their teeth as they do to their complexions, they would be 50 per cent healthier and better looking."
From *Nature* 2 January 1913

1. Lauschke, V. M., Tsiairis, C. D., François, P. & Aulehla, A. *Nature* **493,** 101–105 (2013).
2. Pourquié, O. *Cell* **145,** 650–663 (2011).
3. Oates, A. C., Morelli, L. G. & Ares, S. *Development* **139,** 625–639 (2012).
4. Saga, Y. *Curr. Opin. Genet. Dev.* **22,** 331–338 (2012).
5. Cooke, J. *Nature* **254,** 196–199 (1975).
6. Cooke, J. *Nature* **290,** 775–778 (1981).
7. Cooke, J. & Zeeman, E. C. *J. Theor. Biol.* **58,** 455–476 (1976).
8. Meinhardt, H. *Models of Biological Pattern Formation* (Academic, 1982).
9. Delaune, E. A., François, P., Shih, N. P. & Amacher, S. L. *Dev. Cell* **23,** 995–1005 (2012).
10. Soroldoni, D. & Oates, A. C. *Curr. Opin. Genet. Dev.* **21,** 600–605 (2011).
11. Ben-Zvi, D., Pyrowolakis, G., Barkai, N. & Shilo, B.-Z. *Curr. Biol.* **21,** 1391–1396 (2011).
12. Ben-Zvi, D. & Barkai, N. *Proc. Natl Acad. Sci. USA* **107,** 6924–6929 (2010).
13. Gregor, T., Bialek, W., de Ruyter van Steveninck, R. R., Tank, D . W. & Wieschaus, E. F. *Proc. Natl Acad. Sci. USA* **102,** 18403–18407 (2005).
14. Hamaratoglu, F., de Lachapelle, A. M., Pyrowolakis, G., Bergmann, S. & Affolter, M. *PLoS Biol.* **9,** e1001182 (2011).

STRUCTURAL BIOLOGY

Membrane enzyme cuts a fine figure

Malfunction of presenilin enzymes, which cleave proteins in cell membranes, can lead to Alzheimer's disease. A crystal structure of a microbial presenilin provides insights into the workings of this enzyme family. SEE ARTICLE P.56

MICHAEL S. WOLFE

The interior of a cell membrane is a water-repelling environment. So the discovery[1] that some protease enzymes use water molecules to cut other proteins within membranes was surprising. Three types of such enzymes, which have a variety of roles in biology and disease, have been identified: zinc-containing site-2 proteases, rhomboid serine proteases and aspartyl proteases, such as presenilin. Atomic-resolution structures of site-2 protease and rhomboid enzymes[2,3] have greatly improved our understanding of the mechanisms by which these proteases cleave their substrate proteins, but such a structure for a presenilin has remained elusive until now. On page 56 of this issue, Li and colleagues[4] describe the first detailed structure of a presenilin-type protein, providing a framework for future mechanistic studies and drug-discovery programmes*.

There are two types of presenilin: those that function as single polypeptides, such as the signal peptide peptidase[5], and those that require other proteins for activity. Presenilins of the second type assemble into γ-secretases, which are enzyme complexes, composed of four different proteins, that cleave many single-pass membrane proteins (each one containing a single transmembrane domain, or TMD), including the Notch receptor and the amyloid-β precursor protein (APP)[6]. The cleavage regulates the functions of the target proteins and releases peptides that can have various activities. Functional γ-secretases are essential for Notch signalling processes[7,8], which regulate cell differentiation during development and adulthood in multicellular animals. Moreover, mutations in genes encoding presenilins can cause early-onset Alzheimer's disease by altering how the amyloid-β protein is produced from APP cleavage.

Electron microscopy has provided low-resolution structural images of γ-secretase complexes, but the production of an atomic-resolution crystal structure will be highly challenging. Li *et al.* show that this is much more feasible (although still not easy) for a presenilin alone, representing an important step towards determining the structure of the entire enzyme complex. Their report provides a case study in how to approach the crystallization and structure determination of a membrane protein, an area of investigation currently at the frontier of structural biology.

The authors set out to identify a presenilin, or a presenilin-type protein, that could be overproduced in a bacterial host, purified in its active form and concentrated sufficiently for crystallization trials. After trying several proteins derived from a variety of organisms, they focused on a protease (mmPSH) from the archaean microorganism *Methanoculleus marisnigri*. This effort involved considerable protein engineering, leading to the identification of five mutations that improved the solubility of mmPSH, supported protease function and allowed the formation of high-quality crystals for structure determination.

The structure of mmPSH is only a snapshot, one stable conformation among a continuum of others. Even so, it confirms that presenilins contain nine TMDs and that two aspartate amino-acid residues — located in TMD6 and TMD7, and known to be essential for protease activity[9] — are close to each other and buried in the membrane (Fig. 1). The structure seems to be consistent with previous biochemical studies of presenilins and γ-secretases regarding the arrangement of the TMDs, the water-accessibility of certain residues purportedly near the active site (the part of the enzyme where the cleavage reaction takes place), and the interaction with substrate proteins[6]. But it provides far more detail than previous studies and presents some surprises as well.

One surprise is that the mmPSH protease has a pore that goes through the entire transmembrane region; it may be one route for water to enter for the cleavage reaction. It

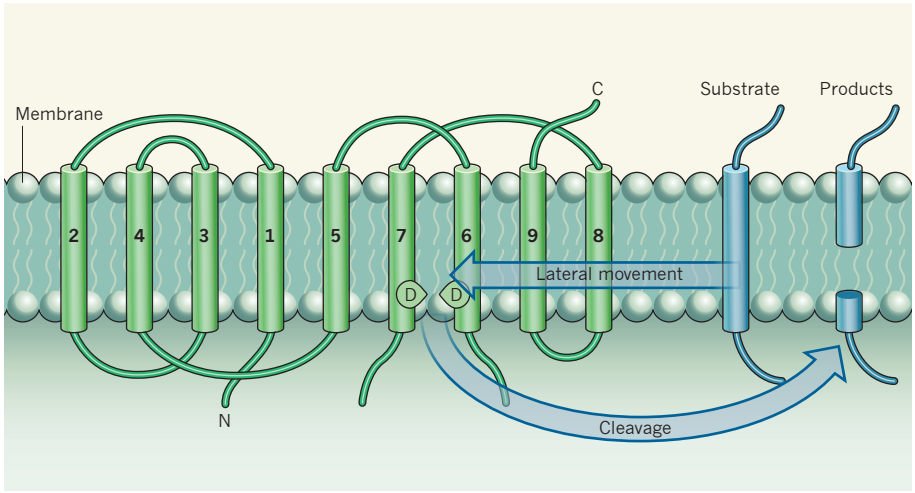


Figure 1 | Architecture of a presenilin enzyme. Presenilins are membrane-embedded protease enzymes that cleave other transmembrane proteins in a regulated manner. Li *et al.*[4] show that a microbial presenilin-type protein, mmPSH, has nine transmembrane domains (TMDs; shown as columns). The active site, which drives the cleavage reaction, is composed of two aspartate amino-acid residues (denoted as D), one in TMD6 and the other in TMD7. The authors' results suggest that lateral movement of the substrate protein into the active site for cleavage is gated by TMD9. The mmPSH protein contains cavities (not shown) that might allow entry of water, which is activated by the catalytic aspartates for substrate cleavage into products. N and C represent the amino and carboxy termini of the protein.

*This article and the paper under discussion[4] were published online on 19 December 2012.

would be interesting to know whether ions or small molecules can traverse this pore, because presenilin has been reported to act as a calcium channel[10]. Alternatively, the pore might be plugged by a lipid or other small molecules in the membrane.

Another surprise is that the protein associates to form tetramers (composed of four mmPSH units), although the functional significance of this finding is unclear. It has been suggested that several units of presenilin might be present in the γ-secretase complex, but such an oligomeric arrangement is not required for catalysis[11]. Nevertheless, a tetrameric organization of signal peptide peptidases and presenilins cannot be ruled out, and the new structure will allow the design of specific experiments to test its relevance.

The mmPSH structure also suggests how a substrate protein might interact with the protease and gain access to the active site. This is an especially complicated issue for proteases embedded in the membrane: the cleavage site is within the substrate's TMD, and the water-containing active site is protected from the water-repellent environment of the membrane. The substrate, which is restricted to movement in two dimensions within the membrane, must therefore first interact with the outer surface of the protease before gaining access to the internal active site. The route of substrate entry in mmPSH is apparently between TMD6 and TMD9, which is consistent with previous research[6] on the interaction of substrates with human γ-secretase.

However, the new crystal structure requires some conformational adjustment for the catalytic aspartates to be properly aligned and close enough to interact with each other and to activate water for substrate cleavage. The authors suggest that interaction with the substrate may lead to correct alignment of the aspartates. This is a realistic possibility, but another is that the protease must be embedded in membranes of appropriate lipid composition or be solubilized by activity-supporting detergents. In any event, the reported structure is only one stable conformation of a presenilin protease, and future crystal structures that capture other conformations might provide a clearer sense of protease dynamics and substrate interaction. This is what happened for rhomboid enzymes[12] and site-2 proteases[3].

Li *et al.* also tried to provide more-specific insights into human presenilin, particularly into how mutations that cause Alzheimer's might affect protease activity. First, they built a structural model of human presenilin, based on the mmPSH structure. Then they created several mmPSH variants containing specific changes in amino-acid residues that mirrored disease-causing mutations, and explained the resultant effects on enzyme activity on the basis of the position of these mutations in their human presenilin model.

Despite reasonable similarity between the two proteins, this is the point at which the limitations of extrapolating from the microbial protease to human presenilin in active protease complexes are reached. The effects of the mutations in mmPSH varied: some reduced or abolished activity, others did not. The relevance of this result is debatable, given that the consistent effect of the mutations in human presenilin is to increase the proportion of longer amyloid-β peptides with a higher tendency to form aggregates (a process associated with Alzheimer's disease), and not a reduction of overall protease activity, although this can occur with some mutations[13,14].

Despite these caveats, the mmPSH structure provides clear insight into the nature and function of membrane-embedded aspartyl proteases. The microbial protein is likely to have the same basic fold and core structure as human presenilins, so Li and colleagues' work opens up a whole new horizon that should ultimately lead to a detailed understanding of human presenilins and of the entire γ-secretase complex. Such understanding should lead to specific ideas about how disease-causing mutations alter function, and how small molecules might be designed for safe and effective treatment of Alzheimer's disease. ■

Michael S. Wolfe *is at the Center for Neurologic Diseases, Harvard Medical School and Brigham and Women's Hospital, Boston, Massachusetts 02115, USA.*
e-mail: mwolfe@rics.bwh.harvard.edu

1. Wolfe, M. S. *Chem. Rev.* **109,** 1599–1612 (2009).
2. Wang, Y., Zhang, Y. & Ha, Y. *Nature* **444,** 179–180 (2006).
3. Feng, L. *et al. Science* **318,** 1608–1612 (2007).
4. Li, X. *et al. Nature* **493,** 56–61 (2013).
5. Weihofen, A., Binns, K., Lemberg, M. K., Ashman, K. & Martoglio, B. *Science* **296,** 2215–2218 (2002).
6. De Strooper, B., Iwatsubo, T. & Wolfe, M. S. *Cold Spring Harb. Perspect. Med.* **2,** a006304 (2012).
7. De Strooper, B. *et al. Nature* **398,** 518–522 (1999).
8. Struhl, G. & Greenwald, I. *Nature* **398,** 522–525 (1999).
9. Wolfe, M. S. *et al. Nature* **398,** 513–517 (1999).
10. Tu, H. *et al. Cell* **126,** 981–993 (2006).
11. Sato, T. *et al. J. Biol. Chem.* **282,** 33985–33993 (2007).
12. Wu, Z. *et al. Nature Struct. Mol. Biol.* **13,** 1084–1091 (2006).
13. Quintero-Monzon, O. *et al. Biochemistry* **50,** 9023–9035 (2011).
14. Chávez-Gutiérrez, L. *et al. EMBO J.* **31,** 2261–2274 (2012).

CLIMATE CHANGE

All in the timing

How influential are the various factors involved in curbing global warming? A study finds that the timing of emissions reduction has the largest impact on the probability of limiting temperature increases to 2 °C. SEE LETTER P.79

STEVE HATFIELD-DODDS

Climate science sometimes seems to have overtaken economics as the most dismal science. But a study by Rogelj *et al.* on page 79 of this issue[1] might just change that. The authors quantify the importance of five 'uncertainties' that are thought to influence the chance of limiting global temperatures to different levels, using a suite of models to generate around 500 scenario variations. They find that the timing of international action to limit emissions has by far the largest impact. Furthermore, the models show that the impact of timing is highly nonlinear, and that delaying emissions limits by only five years, from 2020 to 2025, would dramatically cut the likelihood of limiting warming to 2 °C. The findings should help to make risks and consequences more transparent, and thereby support better-informed economic and political decisions.

The five major uncertainties assessed by Rogelj and colleagues were the following: the responsiveness of the physical climate system to cumulative emissions; the deployment of energy- and land-based emission-reduction technologies; the global demand for energy (which includes combined uncertainties about population, income growth and energy efficiency); the global carbon price that the international community is willing to impose; and the timing of substantive action to limit emissions (phased in from 2010). The analysis covers limiting the temperature in 2100 to 1.5, 2, 2.5 and 3 °C above pre-industrial levels, with a focus on 2 °C.

These scenario comparisons revealed timing of global action to be the uncertainty with the greatest effect. For example, the authors find that bringing forward global action on emissions from 2020 to 2015 would improve the chance of limiting temperatures to 2 °C from 56% to 60%, all else being equal. To put this another way, achieving the same 60% chance of success with action starting in 2020 would require a 2020 carbon price of around US$150 per tonne of carbon dioxide equivalent (CO_2e) — more than double the $60 per tonne CO_2e required if action begins in 2015. However, delaying emissions limits from 2020 to 2025 would bring the chance of success down to 34%, and the authors found no scenario in which a feasible increase in carbon price or improvements in energy technology could make up for these five years of delay.

Geophysical uncertainties are the next most

significant factor, followed by societal factors influencing energy demand, and then uncertainties surrounding mitigation technologies. This information brings clarity to the relative contributions of the timing of global action; the role of energy efficiency; potential mitigation technologies; and required carbon prices. And it will help to assess trade-offs that might be made between these factors. Rogelj and colleagues' analysis also suggests that higher carbon prices will drive more-rapid deployment of low-emission technologies and improve the chance of success, but that imposing a carbon price above $150 would make very little difference to temperature outcomes, because this carbon-price path is sufficient to drive the uptake of all the modelled reduction technologies.

The study, and another recent analysis by some of the same authors[2], also find that the use of different technologies affects both the chance and cost of success. Here, the key results are that the development and deployment of carbon capture and storage technology for fossil fuels and, subsequently, bioenergy are essential for limiting temperature increases to 1.5 °C, and that failure to deploy such technology would reduce the chance of limiting temperature increases to 2 °C by 14–16 percentage points (with a 2020 carbon price of $60–150).

Rogelj and colleagues' assessment complements existing published economic analysis of the costs and benefits of emissions reductions. That literature is characterized by two underappreciated points of consensus. First, the key protagonists in the debate over the British government's 2006 Stern Review[3] on the economics of climate change now all agree that action to limit temperature rise to 2 °C (or emissions to 450 parts per million CO_2e) would provide net benefits[4,5]; this represents a quiet reversal[6,7] in the position of economist William Nordhaus, who was initially critical of the review. Second, economic analysis shows that ambitious global action to limit emissions is fully consistent with strong economic growth and improvements in living standards[8]. A recent multi-model review[9] finds, for example, that average global income is projected to roughly double over the four decades to 2050 across all scenarios targeting 450 parts per million CO_2e, and that the average income growth in emerging economies and other developing countries would be above the global average.

Perhaps the only significant limitation of Rogelj and colleagues' analysis is that the exploration of geophysical uncertainties does not fully incorporate known positive feedbacks on climate change, particularly the effect of CO_2 and methane releases from warming permafrost[10,11]. (This is a limitation of climate modelling generally.) Although the timing and magnitude of these releases are highly uncertain, on current temperature trends the cumulative emissions from permafrost are considered likely to be more than 30 gigatonnes CO_2e by 2040, with significant emissions continuing for more than two centuries[11,12]. This has the potential to dwarf reductions in anthropogenic emissions. However, this consideration simply strengthens the case for early and decisive action to limit anthropogenic emissions, because it would reduce the temperature increases that could trigger such feedbacks. This suggests that it would be valuable for economic studies to explore the 'insurance value' of reducing the risk of climate feedbacks.

Minor limitations of the study are that it uses a single modelling suite for analysis and is based on a single, somewhat optimistic view of underlying population and economic trends[13,14]. This implies that the authors' estimated probabilities of achieving particular temperature limits may also be somewhat optimistic, and that it would be good to replicate this work using other models and a wider range of underlying scenarios.

Rogelj *et al.* have provided a new benchmark for assessing the relative contributions of several major uncertainties in the quest to limit climate change. The study's key message reinforces previous findings[14–16] that urgent and more ambitious global action is required to maintain any chance of limiting global warming to 2 °C. The clear finding that the world would be better off acting from 2015 rather than 2020 also raises sharp and serious questions about the trade-offs implicit in the current pace of global negotiations and action. The window for effective action on climate change is closing quickly, and Rogelj *et al.* have put a price tag on each year of delay. ■

Steve Hatfield-Dodds *is at CSIRO Ecosystem Sciences, Black Mountain Laboratories, Canberra, ACT 2601, Australia, and at the Crawford School of Public Policy, Australian National University, Canberra.*
e-mail: steve.hatfield-dodds@csiro.au

1. Rogelj, J., McCollum, D. L., Reisinger A., Meinshausen, M. & Riahi, K. *Nature* **493,** 79–83 (2013).
2. Rogelj, J., McCollum, D. L., O'Neill, B. C. & Riahi, K. *Nature Clim. Change* http://dx.doi.org/10.1038/nclimate1758 (2012).
3. Stern, N. *The Economics of Climate Change: The Stern Review* (HM Treasury, 2006).
4. Stern, N. *Am. Econ. Rev.* **98**(2), 1–37 (2008).
5. Heal, G. *Rev. Environ. Econ. Policy* **3,** 4–21 (2009).
6. Nordhaus, W. *A Question of Balance: Weighing the Options on Global Warming Policies* (Yale Univ. Press, 2008).
7. Nordhaus, W. D. *Proc. Natl Acad. Sci. USA* **107,** 11721–11726 (2010).
8. Garnaut, R. *The Garnaut Review 2011: Australia in the Global Response to Climate Change* (Cambridge Univ. Press, 2011).
9. Klinsky, S., Hatfield-Dodds, S. & Mizuno, E. *Living Standards and Economic Performance with Ambitious Climate Action* (Climate Strategies, 2012).
10. Lenton, T. M. *et al. Proc. Natl Acad. Sci. USA* **105,** 1786–1793 (2008).
11. Schuur, E. A. G., Abbot, B. & the Permafrost Carbon Network *Nature* **480,** 32–33 (2011).
12. Schaefer, K., Lantuit, H., Romanovsky, V. E., Schuur, E. A. G. & Gärtner-Roer, I. *Policy Implications of Warming Permafrost* (UNEP, 2012).
13. Garnaut, R., Howes, S., Jotzo, F. & Sheehan, P. *Oxford Rev. Econ. Policy* **24,** 377–401 (2008).
14. Peters, G. P. *et al. Nature Clim. Change* http://dx.doi.org/10.1038/nclimate1783 (2012).
15. Rogelj, J. *et al. Nature* **464,** 1126–1128 (2010).
16. United Nations Environment Programme. *The Emissions Gap Report 2012: A UNEP Synthesis Report* (UNEP, 2012).

MICROBIOLOGY

Break down the walls

Nanoscale imaging reveals that bacterial and fungal enzymes use different mechanisms to deconstruct plant cell walls. The finding may provide clues about how to enhance the efficiency of liquid-biofuel production from biomass.

RICHARD A. DIXON

Plants are increasingly being used as raw materials in the production of ethanol and other liquid biofuels. But the poor accessibility of sugars embedded in plant cell walls — known as recalcitrance — is a major barrier to economically viable implementation of these technologies[1]. Although recalcitrance is an inherent property of plant cell walls, different microorganisms use different enzymes to degrade the walls, and a lack of understanding of these interactions has limited the design of plants that have reduced recalcitrance. Writing in *Science*, Ding *et al.*[2] employ state-of-the-art microscopy techniques to show, at nanometre resolution, plant cell walls being degraded through distinct mechanisms by bacteria and fungi.

The saying that there is no such thing as a free lunch seems particularly apposite when considering the evolution of plant cell walls. Some plant cells are surrounded by a secondary cell wall that includes a thick layer of cellulose (a polysaccharide) associated with the complex organic polymer lignin. This secondary thickening provides physical strength to support upright growth, hydrophobicity to allow water transport, and protection against microbial ingress. In parallel, microbes have evolved mechanisms, including cellulose-degrading enzymes, to degrade plant cell walls to access the plants' nutritious sugars. Fungi use degradative enzymes called cellulases,

significant factor, followed by societal factors influencing energy demand, and then uncertainties surrounding mitigation technologies. This information brings clarity to the relative contributions of the timing of global action; the role of energy efficiency; potential mitigation technologies; and required carbon prices. And it will help to assess trade-offs that might be made between these factors. Rogelj and colleagues' analysis also suggests that higher carbon prices will drive more-rapid deployment of low-emission technologies and improve the chance of success, but that imposing a carbon price above $150 would make very little difference to temperature outcomes, because this carbon-price path is sufficient to drive the uptake of all the modelled reduction technologies.

The study, and another recent analysis by some of the same authors[2], also find that the use of different technologies affects both the chance and cost of success. Here, the key results are that the development and deployment of carbon capture and storage technology for fossil fuels and, subsequently, bioenergy are essential for limiting temperature increases to 1.5 °C, and that failure to deploy such technology would reduce the chance of limiting temperature increases to 2 °C by 14–16 percentage points (with a 2020 carbon price of $60–150).

Rogelj and colleagues' assessment complements existing published economic analysis of the costs and benefits of emissions reductions. That literature is characterized by two underappreciated points of consensus. First, the key protagonists in the debate over the British government's 2006 Stern Review[3] on the economics of climate change now all agree that action to limit temperature rise to 2 °C (or emissions to 450 parts per million CO_2e) would provide net benefits[4,5]; this represents a quiet reversal[6,7] in the position of economist William Nordhaus, who was initially critical of the review. Second, economic analysis shows that ambitious global action to limit emissions is fully consistent with strong economic growth and improvements in living standards[8]. A recent multi-model review[9] finds, for example, that average global income is projected to roughly double over the four decades to 2050 across all scenarios targeting 450 parts per million CO_2e, and that the average income growth in emerging economies and other developing countries would be above the global average.

Perhaps the only significant limitation of Rogelj and colleagues' analysis is that the exploration of geophysical uncertainties does not fully incorporate known positive feedbacks on climate change, particularly the effect of CO_2 and methane releases from warming permafrost[10,11]. (This is a limitation of climate modelling generally.) Although the timing and magnitude of these releases are highly uncertain, on current temperature trends the cumulative emissions from permafrost are considered likely to be more than 30 gigatonnes CO_2e by 2040, with significant emissions continuing for more than two centuries[11,12]. This has the potential to dwarf reductions in anthropogenic emissions. However, this consideration simply strengthens the case for early and decisive action to limit anthropogenic emissions, because it would reduce the temperature increases that could trigger such feedbacks. This suggests that it would be valuable for economic studies to explore the 'insurance value' of reducing the risk of climate feedbacks.

Minor limitations of the study are that it uses a single modelling suite for analysis and is based on a single, somewhat optimistic view of underlying population and economic trends[13,14]. This implies that the authors' estimated probabilities of achieving particular temperature limits may also be somewhat optimistic, and that it would be good to replicate this work using other models and a wider range of underlying scenarios.

Rogelj *et al.* have provided a new benchmark for assessing the relative contributions of several major uncertainties in the quest to limit climate change. The study's key message reinforces previous findings[14–16] that urgent and more ambitious global action is required to maintain any chance of limiting global warming to 2 °C. The clear finding that the world would be better off acting from 2015 rather than 2020 also raises sharp and serious questions about the trade-offs implicit in the current pace of global negotiations and action. The window for effective action on climate change is closing quickly, and Rogelj *et al.* have put a price tag on each year of delay. ■

Steve Hatfield-Dodds *is at CSIRO Ecosystem Sciences, Black Mountain Laboratories, Canberra, ACT 2601, Australia, and at the Crawford School of Public Policy, Australian National University, Canberra.*
e-mail: steve.hatfield-dodds@csiro.au

1. Rogelj, J., McCollum, D. L., Reisinger A., Meinshausen, M. & Riahi, K. *Nature* **493,** 79–83 (2013).
2. Rogelj, J., McCollum, D. L., O'Neill, B. C. & Riahi, K. *Nature Clim. Change* http://dx.doi.org/10.1038/nclimate1758 (2012).
3. Stern, N. *The Economics of Climate Change: The Stern Review* (HM Treasury, 2006).
4. Stern, N. *Am. Econ. Rev.* **98**(2), 1–37 (2008).
5. Heal, G. *Rev. Environ. Econ. Policy* **3,** 4–21 (2009).
6. Nordhaus, W. *A Question of Balance: Weighing the Options on Global Warming Policies* (Yale Univ. Press, 2008).
7. Nordhaus, W. D. *Proc. Natl Acad. Sci. USA* **107,** 11721–11726 (2010).
8. Garnaut, R. *The Garnaut Review 2011: Australia in the Global Response to Climate Change* (Cambridge Univ. Press, 2011).
9. Klinsky, S., Hatfield-Dodds, S. & Mizuno, E. *Living Standards and Economic Performance with Ambitious Climate Action* (Climate Strategies, 2012).
10. Lenton, T. M. *et al. Proc. Natl Acad. Sci. USA* **105,** 1786–1793 (2008).
11. Schuur, E. A. G., Abbot, B. & the Permafrost Carbon Network *Nature* **480,** 32–33 (2011).
12. Schaefer, K., Lantuit, H., Romanovsky, V. E., Schuur, E. A. G. & Gärtner-Roer, I. *Policy Implications of Warming Permafrost* (UNEP, 2012).
13. Garnaut, R., Howes, S., Jotzo, F. & Sheehan, P. *Oxford Rev. Econ. Policy* **24,** 377–401 (2008).
14. Peters, G. P. *et al. Nature Clim. Change* http://dx.doi.org/10.1038/nclimate1783 (2012).
15. Rogelj, J. *et al. Nature* **464,** 1126–1128 (2010).
16. United Nations Environment Programme. *The Emissions Gap Report 2012: A UNEP Synthesis Report* (UNEP, 2012).

MICROBIOLOGY

Break down the walls

Nanoscale imaging reveals that bacterial and fungal enzymes use different mechanisms to deconstruct plant cell walls. The finding may provide clues about how to enhance the efficiency of liquid-biofuel production from biomass.

RICHARD A. DIXON

Plants are increasingly being used as raw materials in the production of ethanol and other liquid biofuels. But the poor accessibility of sugars embedded in plant cell walls — known as recalcitrance — is a major barrier to economically viable implementation of these technologies[1]. Although recalcitrance is an inherent property of plant cell walls, different microorganisms use different enzymes to degrade the walls, and a lack of understanding of these interactions has limited the design of plants that have reduced recalcitrance. Writing in *Science*, Ding *et al.*[2] employ state-of-the-art microscopy techniques to show, at nanometre resolution, plant cell walls being degraded through distinct mechanisms by bacteria and fungi.

The saying that there is no such thing as a free lunch seems particularly apposite when considering the evolution of plant cell walls. Some plant cells are surrounded by a secondary cell wall that includes a thick layer of cellulose (a polysaccharide) associated with the complex organic polymer lignin. This secondary thickening provides physical strength to support upright growth, hydrophobicity to allow water transport, and protection against microbial ingress. In parallel, microbes have evolved mechanisms, including cellulose-degrading enzymes, to degrade plant cell walls to access the plants' nutritious sugars. Fungi use degradative enzymes called cellulases,

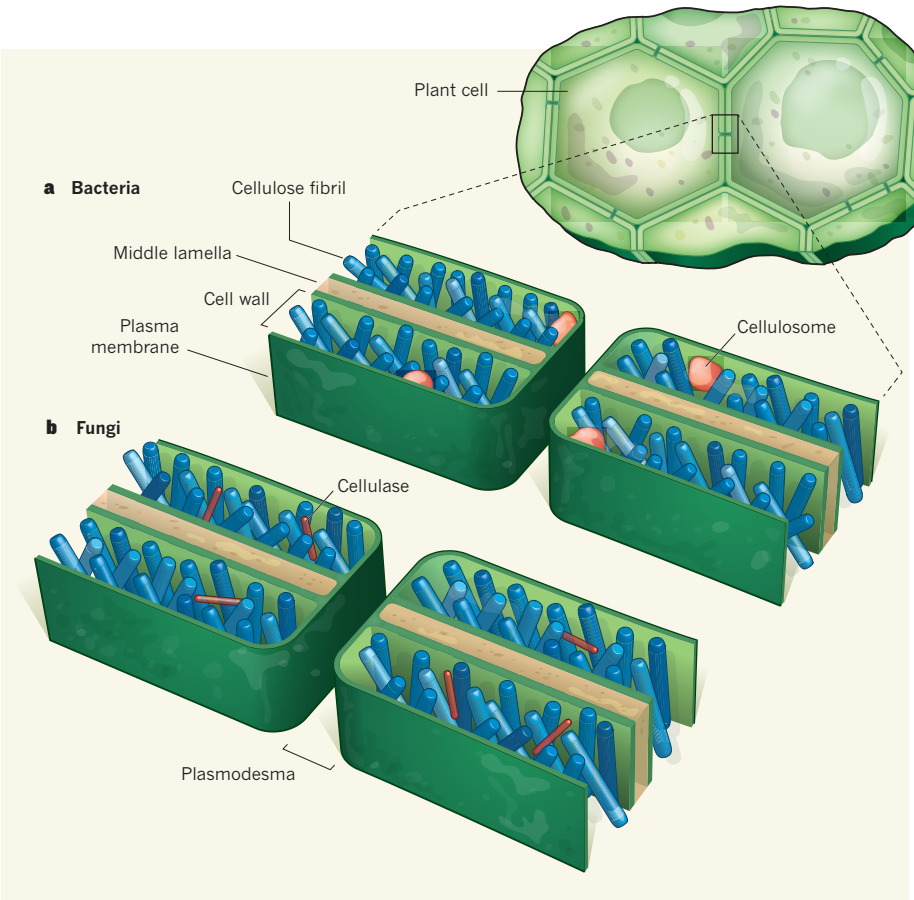


Figure 1 | Degradation of plant cell walls by microbial enzymes. Cellulolytic enzymes degrade cellulose, the polymer that forms a significant portion of the cell walls of plants. Many bacteria produce complexes of cellulolytic enzymes called cellulosomes, whereas fungi attack plant cell walls by secreting free cellulolytic enzymes called cellulases. Ding *et al.*[2] used nanometre-scale microscopy to show that plant degradation by these two strategies occurs through different mechanisms. **a**, They observe that bacterial cellulosomes peel off individual cellulose microfibrils from the cell-wall surface and around plasmodesmata (small tubes that connect adjacent plant cells). **b**, By contrast, their microscope images show fungal cellulases penetrating deep into the cellulose microfibrillar network.

whereas in bacteria, multiple enzymes self-assemble into a complex called the cellulosome[3]. However, the lignin in secondarily thickened plant cell walls presents an effective barrier to most microbial enzyme systems. Many of the feedstock materials used for biofuel production are lignocellulosic, including perennial grasses, trees and maize (corn) stover — the leaves and stalks of the plant left after grain harvest.

In the United States, the 2005 Energy Policy Act included a Renewable Fuel Standard (RFS), which gives a minimum volume of biofuels to be used in the national transportation fuel supply each year. However, the US National Academy of Sciences recently concluded[4] that in the absence of "major technological innovation … the RFS2-mandated consumption of 16 billion gallons of ethanol-equivalent cellulosic biofuels by 2022 is unlikely to be met". Overcoming cell-wall recalcitrance may be the technological innovation that has the largest impact on reducing production costs of liquid biofuels from lignocellulosic feedstocks. It is clear that reducing lignin content, either by taking advantage of natural variation in plant varieties or through targeted transgenic approaches, can enhance the release of cell-wall sugars[5,6], although this can have negative effects on plant growth[7]. But might there be other ways to target recalcitrance? Ding and colleagues sought to pinpoint which features of the plant cell wall most affect their microbial digestibility.

The authors used a suite of correlated, multimodal microscopy approaches to observe, in real time, fluorescently labelled microbial enzymes degrading cell-wall polysaccharides of maize stover. This revealed intriguing differences between the digestive strategies of bacteria and fungi (Fig. 1). The images show that bacterial cellulosomes start digesting the cell walls away from the middle lamella — the cell-wall layer that joins adjacent plant cells. By contrast, fungal cellulases dissolve the wall in a uniform manner from the innermost side, leaving the middle lamella intact. Furthermore, it seems that fungal cellulases penetrate the cellulose microfibrillar network to produce digestion pits, whereas cellulosomes peel off individual microfibrils from the cell-wall surface.

The authors' observations also confirm that lignin is indeed the major impediment to cell-wall degradation by either fungal cellulases or bacterial cellulosomes. But they take the story an important step forward by showing that the main problem is that lignin physically masks the cellulose microfibrils, rather than absorbing lignin-degrading enzymes. By quantifying images obtained by atomic force microscopy, the authors show that the hydrophobic planar faces of cellulose molecules are the preferred binding sites of microbial enzymes, and are therefore crucial for enzyme access. The hydrophobic faces are exposed in primary cell walls, but are masked by lignin in secondary walls, either partially in certain parenchyma cells or totally in schlerenchyma-fibre cells.

What do these findings mean for the development of biomass feedstocks? Lignification of some cell types occurs after the plant has finished growing, but early harvesting of perennial crops presents sustainability problems because essential nutrients such as nitrogen and phosphorus are transferred back to the root system only during this senescent phase of the plant life cycle[8]. Ding and colleagues' study suggests that it might be just as effective, if not more so, to engineer plants to have reduced levels of lignin–polysaccharide linkages as it would be to attempt to reduce lignin levels per se. This would produce a cellulose-microfibril structure that is more conducive to degradation and would substantially reduce the severity of pretreatment required to render lignocellulosic biomass suitable for biofuel generation.

It will be interesting to apply the microscopy techniques used by Ding and colleagues to probe the interactions between microbial enzyme systems and plants that have been genetically modified to have reduced recalcitrance through targeting of cell-wall components. Only by understanding recalcitrance from the perspectives of both the plant and the microbe can more effective bioprocessing systems be developed. However, whether modifying plant cell walls for optimal deconstruction will allow microorganisms to jump the lunch queue and cause disease in their plant targets remains to be seen. ■

Richard A. Dixon *is in the Department of Biological Sciences, University of North Texas, Denton, Texas 76203, USA.*
e-mail: richard.dixon@unt.edu

1. Himmel, M. E. *et al. Science* **315,** 804–807 (2007).
2. Ding, S.-Y. *et al. Science* **338,** 1055–1060 (2012).
3. Bomble, Y. J. *et al. Biol. Chem.* **286,** 5614–5623 (2011).
4. National Academy of Sciences. www.nap.edu/openbook.php?record_id=13105&page=R1 (2011).
5. Studer, M. H. *et al. Proc. Natl Acad. Sci. USA* **108,** 6300–6305 (2011).
6. Chapple, C., Ladisch, M. & Meilan, R. *Nature Biotechnol.* **25,** 746–748 (2007).
7. Hoffmann, L. *et al. Plant Cell* **16,** 1446–1465 (2004).
8. Yang, J. *Bioenergy Res.* **2,** 257–266 (2009).

ARTICLE

doi:10.1038/nature11732

Non-Fermi-liquid *d*-wave metal phase of strongly interacting electrons

Hong-Chen Jiang[1], Matthew S. Block[2], Ryan V. Mishmash[3], James R. Garrison[3], D. N. Sheng[4], Olexei I. Motrunich[5] & Matthew P. A. Fisher[3]

Developing a theoretical framework for conducting electronic fluids qualitatively distinct from those described by Landau's Fermi-liquid theory is of central importance to many outstanding problems in condensed matter physics. One such problem is that, above the transition temperature and near optimal doping, high-transition-temperature copper-oxide superconductors exhibit 'strange metal' behaviour that is inconsistent with being a traditional Landau Fermi liquid. Indeed, a microscopic theory of a strange-metal quantum phase could shed new light on the interesting low-temperature behaviour in the pseudogap regime and on the *d*-wave superconductor itself. Here we present a theory for a specific example of a strange metal—the '*d*-wave metal'. Using variational wavefunctions, gauge theoretic arguments, and ultimately large-scale density matrix renormalization group calculations, we show that this remarkable quantum phase is the ground state of a reasonable microscopic Hamiltonian—the usual *t*–*J* model with electron kinetic energy *t* and two-spin exchange *J* supplemented with a frustrated electron 'ring-exchange' term, which we here examine extensively on the square lattice two-leg ladder. These findings constitute an explicit theoretical example of a genuine non-Fermi-liquid metal existing as the ground state of a realistic model.

Over the past several decades, experiments on strongly correlated materials have routinely revealed, in certain parts of the phase diagram, conducting liquids with physical properties that are qualitatively inconsistent with Landau's Fermi-liquid theory[1]. Examples of these 'non-Fermi-liquid' metals[2] include the strange-metal phase of the copper-oxide superconductors[3,4] and the heavy fermion materials near a quantum critical point[5,6]. However, such non-Fermi-liquid behaviour has been challenging to characterize theoretically, largely owing to the lack of a weakly interacting quasiparticle description. It is even difficult to define a non-Fermi liquid unambiguously, although possible deviations from Fermi-liquid theory include, for example, violation of Luttinger's volume theorem[7], vanishing quasiparticle weight, and anomalous thermodynamics and transport[5,8–12]. This theoretical difficulty is probably preventing a full understanding of the mechanism behind high-temperature superconductivity and also hampering theoretically guided searches for new exotic materials.

Pioneering early theoretical work on the copper oxides relied on two main premises[3,13–17], which guide but do not constrain our pursuit of non-Fermi-liquid physics: (1) that the microscopic behaviour can be described by the square lattice Hubbard model with on-site Coulomb repulsion, which at strong coupling reduces in its simplest form to the *t*–*J* model; and (2) that the physics of the system can be faithfully represented by the 'slave-boson' technique, in which the physical electron operator is written as the product of a slave boson ('chargon'), which carries the electronic charge, and a spin-half fermionic 'spinon'[18], which carries the spin (both chargon and spinon are strongly coupled to an emergent gauge field). However, within the slave-boson formulation, it has been difficult to access non-Fermi-liquid physics at low temperatures because this requires the chargons to be in an uncondensed, yet conducting, quantum phase[19]—the elusive 'Bose metal'. Early attempts to describe the strange metal in this framework treated it as a strictly finite-temperature phenomenon in which the slave bosons form an uncondensed, but classical, Bose fluid[15,16], a treatment which excludes the possibility that the strange metal is a true quantum phase at all.

In our view, the strange metal should be viewed as a genuine two-dimensional (2D) quantum phase, which may be unstable to superconducting or pseudogap behaviour. Indeed, recent experimental work on $La_{2-x}Sr_xCuO_4$ has shown that when superconductivity is stripped away by high magnetic fields, strange-metal behaviour persists over a wide doping range down to extremely low temperatures[20]. Thus, the strange metal in the copper oxides is quite possibly a true, extended, zero-temperature quantum phase[4].

Inspired by these results and building on our previous work, which proposed[21] and realized[22–24] a true, zero-temperature Bose metal, we use a variant of the slave-boson approach to construct and analyse an exotic 2D non-Fermi-liquid quantum phase, which we refer to as the '*d*-wave metal'. The *d*-wave metal is modelled by a variational wavefunction consisting of a product of a *d*-wave Bose-metal wavefunction[21–24] for the chargons and a usual Slater determinant for the spinons. Importantly, placing the chargons in the *d*-wave Bose metal state provides the many-electron wavefunction with a sign structure that is qualitatively distinct from that of a simple Slater determinant, and in particular, imprints strong singlet *d*-wave two-particle correlations. This results in a gapless, conducting quantum fluid with an electron momentum distribution function that exhibits a critical, singular surface that violates Luttinger's volume theorem[7], as well as prominent critical Cooper pairs with *d*-wave character. The *d*-wave nature of our phase is tantalizingly suggestive of incipient *d*-wave superconductivity and is thus of possible relevance to the copper oxides.

Furthermore, tying back into premise (1) above, we propose a reasonably simple model Hamiltonian to stabilize the *d*-wave metal by augmenting the traditional *t*–*J* model with a four-site ring-exchange

[1]Kavli Institute for Theoretical Physics, University of California, Santa Barbara, California 93106, USA. [2]Department of Physics and Astronomy, University of Kentucky, Lexington, Kentucky 40506, USA. [3]Department of Physics, University of California, Santa Barbara, California 93106, USA. [4]Department of Physics and Astronomy, California State University, Northridge, California 91330, USA. [5]Department of Physics, California Institute of Technology, Pasadena, California 91125, USA.

term K. Then, thanks to the numerical and analytical tractability provided by the density matrix renormalization group (DMRG)[25,26] and bosonization[27–30], we can place the problem on a quasi-one-dimensional (1D) two-leg ladder geometry (see Fig. 1). In this system, we establish several lines of compelling evidence that the d-wave metal phase exists as the quantum ground state of our t–J–K model Hamiltonian, and we are able to characterize and understand the phase very thoroughly. Importantly, our realized two-leg d-wave metal state is non-perturbative, in that it cannot be understood within conventional Luttinger liquid theory[27] starting from free electrons[31]. We believe this study to be one of the first unbiased numerical demonstrations of a non-Fermi-liquid metal as the stable ground state of a local Hamiltonian. We also discuss straightforward extensions of these results to two dimensions, and comment on their potential relevance to the actual non-Fermi liquids observed in experiments.

Gauge theory and variational wavefunctions

Our theoretical description of the non-Fermi-liquid d-wave metal begins by writing the electron operator for site $\boldsymbol{r}$ and spin state $s=\uparrow,\downarrow$ as the product of a bosonic chargon $b(\boldsymbol{r})$ and fermionic spinon $f_s(\boldsymbol{r})$; that is, $c_s(\boldsymbol{r})=b(\boldsymbol{r})f_s(\boldsymbol{r})$. With $b(\boldsymbol{r})$ a hard-core boson operator, this construction prohibits doubly occupied sites, an assumption we make from here on. The physical electron Hilbert space is recovered by implementing at each site the constraint $b^\dagger(\boldsymbol{r})b(\boldsymbol{r})=\sum_s f_s^\dagger(\boldsymbol{r})f_s(\boldsymbol{r})=\sum_s c_s^\dagger(\boldsymbol{r})c_s(\boldsymbol{r})=n_e(\boldsymbol{r})$, which physically means that a given site is either empty or contains a chargon and exactly one spinon to compose an electron. Theoretically, this is achieved by strongly coupling the b and f fields via an emergent gauge field[3].

Under the natural assumption that the spinons are in a Fermi-sea state, the behaviour of the chargons determines the resulting electronic phase. Condensing the bosonic chargons so that $\langle b(\boldsymbol{r})\rangle\neq 0$ implies $c_s(\boldsymbol{r})\propto f_s(\boldsymbol{r})$; so, in this case, the electronic phase is that of a Fermi liquid. It then follows that to describe a non-Fermi-liquid conducting quantum fluid within this framework, the chargons must not condense, $\langle b(\boldsymbol{r})\rangle=0$, but must still conduct. However, accessing such a 'Bose metal' phase has proved extremely difficult. In recent work[21–24], however, we have realized a concrete, genuine Bose-metal phase, which we named the "d-wave Bose liquid" or, equivalently, the "d-wave Bose metal" (DBM). The DBM is central to our construction of the d-wave metal. Specifically, in the DBM, we decompose the hard-core boson as $b(\boldsymbol{r})=d_1(\boldsymbol{r})d_2(\boldsymbol{r})$ with the constraint $d_1^\dagger(\boldsymbol{r})d_1(\boldsymbol{r})=d_2^\dagger(\boldsymbol{r})d_2(\boldsymbol{r})=b^\dagger(\boldsymbol{r})b(\boldsymbol{r})$, where d_1 and d_2 are fermionic slave particles ('partons') with anisotropic hopping patterns: d_1 (d_2) is chosen to hop preferentially in the $\hat{x}$ ($\hat{y}$) direction. The resulting bosonic phase is a conducting, yet uncondensed, quantum fluid, which is precisely the phase into which we place the charge sector of the d-wave metal. That is, for the d-wave metal we take an all-fermionic decomposition of the electron

$$c_s(\boldsymbol{r})=d_1(\boldsymbol{r})d_2(\boldsymbol{r})f_s(\boldsymbol{r}) \tag{1}$$

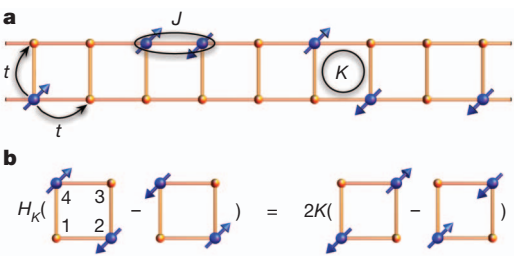


Figure 1 | Schematic of the t–J–K model Hamiltonian. **a**, Picture of the full t–J–K model, equation (6), on the two-leg ladder. We use periodic boundary conditions in the long ($\hat{x}$) direction for all calculations. **b**, Action of the ring term H_K, equation (8), on a single plaquette, elucidating its 'singlet-rotation' nature.

subject to the constraint

$$d_1^\dagger(\boldsymbol{r})d_1(\boldsymbol{r})=d_2^\dagger(\boldsymbol{r})d_2(\boldsymbol{r})=\sum_s f_s^\dagger(\boldsymbol{r})f_s(\boldsymbol{r})=n_e(\boldsymbol{r}) \tag{2}$$

The resulting theory now includes two gauge fields: one to glue together d_1 and d_2 to form the chargon and another to glue together b and f to form the electron. In the Supplementary Information, we give a detailed bosonization analysis of this gauge theory for the two-leg ladder study (see below).

Guided by the slave-boson construction, one can naturally construct electronic variational wavefunctions by taking the product of a hard-core bosonic wavefunction ψ_b and a fermionic wave function ψ_f and evaluating them at the same coordinates (Gutzwiller projection):

$$\psi_c\left(\left\{\boldsymbol{r}_i^\uparrow\right\},\left\{\boldsymbol{r}_i^\downarrow\right\}\right)=\mathcal{P}_G\left[\psi_b(\{\boldsymbol{R}_i\})\times\psi_f\left(\left\{\boldsymbol{r}_i^\uparrow\right\},\left\{\boldsymbol{r}_i^\downarrow\right\}\right)\right] \tag{3}$$

where $\mathcal{P}_G$ performs the projection into the physical electronic Hilbert space: $\{\boldsymbol{R}_i\}=\left\{\boldsymbol{r}_i^\uparrow\right\}\bigcup\left\{\boldsymbol{r}_i^\downarrow\right\}$. If we put the f partons into a spin-singlet Fermi-sea state with orbitals $\{\boldsymbol{k}_j\}$ (Slater determinant), that is, $\psi_f\left(\left\{\boldsymbol{r}_i^\uparrow\right\},\left\{\boldsymbol{r}_i^\downarrow\right\}\right)=\det\left[e^{i\boldsymbol{k}_j\cdot\boldsymbol{r}_i^\uparrow}\right]\det\left[e^{i\boldsymbol{k}_j\cdot\boldsymbol{r}_i^\downarrow}\right]=\psi_f^{FS}$, then we can model both the Fermi-liquid metal and the non-Fermi-liquid d-wave metal in a unified way. In both cases, the wavefunctions are straightforward to implement using variational Monte Carlo (VMC) methods[32–34].

For the Fermi liquid, we put the b partons into a superfluid wavefunction ψ_b^{SF} via a typical Jastrow form, so that, schematically, $\psi_c^{FL}=\mathcal{P}_G\left[\psi_b^{SF}\times\psi_f^{FS}\right]$. Given that ψ_b^{SF} is a positive wavefunction, the sign structure[35] of ψ_c^{FL} is identical to that of the non-interacting Fermi-sea state. In contrast, to model the d-wave metal, we put the b partons into a Bose-metal wavefunction according to the DBM construction of refs 21–24:

$$\psi_b(\{\boldsymbol{R}_i\})=\psi_{d_1}(\{\boldsymbol{R}_i\})\times\psi_{d_2}(\{\boldsymbol{R}_i\})=\psi_b^{DBM} \tag{4}$$

where $\psi_{d_1}\left(\psi_{d_2}\right)$ is a Slater determinant with a Fermi sea compressed in the $\hat{x}$ ($\hat{y}$) direction[21]. Then, we have

$$\psi_c^{d\text{-wave metal}}=\mathcal{P}_G\left[\psi_b^{DBM}\times\psi_f^{FS}\right]=\mathcal{P}_G\left[\psi_{d_1}\times\psi_{d_2}\times\psi_f^{FS}\right] \tag{5}$$

Interestingly, this construction, equation (5), is actually a time-reversal invariant analogue of the composite Fermi-liquid description of the half-filled Landau level[36], where the d-wave Bose-metal wavefunction[21] has the role of Laughlin's $\nu=1/2$ bosonic state[37]. Just as Laughlin's wavefunction imprints a nontrivial complex phase pattern on the Slater determinant, the DBM wavefunction imprints a non-trivial d-wave sign structure. There are many physical signatures associated with putting the chargons into the DBM phase, making the d-wave metal dramatically distinguishable from the traditional Landau Fermi liquid (see below).

Microscopic ring-exchange model

The t–J–K model Hamiltonian which we propose to stabilize the d-wave metal phase is given by

$$H=H_{tJ}+H_K \tag{6}$$

$$H_{tJ}=-t\sum_{\langle i,j\rangle,s=\uparrow,\downarrow}\left(c_{is}^\dagger c_{js}+c_{js}^\dagger c_{is}\right)+J\sum_{\langle i,j\rangle}\boldsymbol{S}_i\cdot\boldsymbol{S}_j \tag{7}$$

$$H_K=2K\sum_{\square}\left(\mathcal{S}_{13}^\dagger\mathcal{S}_{24}+\mathcal{S}_{24}^\dagger\mathcal{S}_{13}\right) \tag{8}$$

where $\langle i,j\rangle$ and $\square$ indicate sums over all nearest-neighbour bonds and all elementary plaquettes of the 2D square lattice, respectively. In the

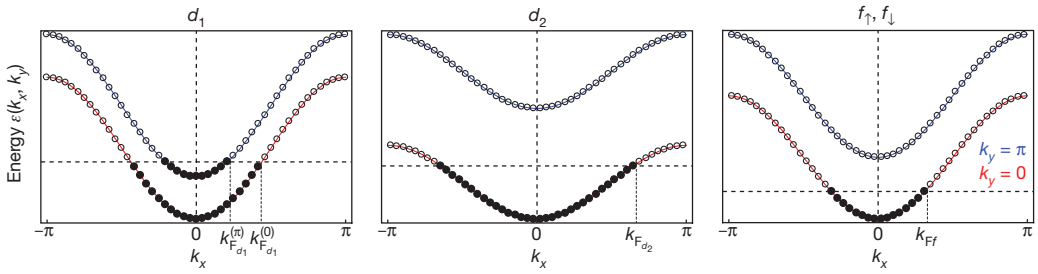


Figure 2 | Picture of the parton bands for the *d*-wave metal phase. We show orbitals for a 48×2 system, showing partially occupied bonding ($k_y = 0$) and antibonding ($k_y = \pi$) bands for d_1 and partially occupied bonding bands for d_2 and $f_{\uparrow/\downarrow}$; that is, each Slater determinant in equation (5) consists of momentum-space orbitals as depicted here. The total electron number is $N_e = N_{c\uparrow} + N_{c\downarrow} = N_{d_1} = N_{d_2} = N_{f\uparrow} + N_{f\downarrow} = 32$, with $N_{f\uparrow} = N_{f\downarrow} = 16$ so that $S_{tot} = 0$; the longitudinal boundary conditions are periodic for d_1 and antiperiodic for d_2 and $f_{\uparrow/\downarrow}$. This is precisely the same *d*-wave metal configuration for which we display characteristic measurements in Fig. 5.

spirit of the *t*–*J* model, we choose to work in the subspace of no doubly occupied sites, but for simplicity, we do ignore the term $-\frac{J}{4} n_i n_j$ present in typical definitions of the *t*–*J* model[3]. In equation (8), we have defined a singlet creation operator on two sites as $\mathcal{S}_{ij}^\dagger = \frac{1}{\sqrt{2}}\left(c_{i\uparrow}^\dagger c_{j\downarrow}^\dagger - c_{i\downarrow}^\dagger c_{j\uparrow}^\dagger\right)$, so that H_K can be viewed as a four-site singlet-rotation term (see Fig. 1). For $K > 0$, the ground state of H_K on a single plaquette with two electrons is a d_{xy}-orbital spin-singlet; so, loosely speaking, H_K has a tendency to build *d*-wave correlations into the system and qualitatively alter the sign structure of the electronic ground state. Further arguments for studying this model in our investigation of the *d*-wave metal can be found in the Supplementary Information.

Although not particularly conventional, our ring-exchange term H_K (which should not be confused with four-site cyclic spin-exchange[38–40]) is present when projecting the continuum many-body Hamiltonian for screened Coulomb-interacting electrons into a narrow, tight-binding band[41] (see Supplementary Information). In fact, estimating the strength of *K*, or coefficients on related terms, in real materials such as $La_{2-x}Sr_xCuO_4$ is an interesting unresolved question.

DMRG and VMC study of two-leg model

Unfortunately, as with any interacting fermionic model, our *t*–*J*–*K* Hamiltonian suffers from the 'fermionic sign problem', rendering quantum Monte Carlo calculations inapplicable[42]. We thus follow the heretofore successful[22–24,39,40] approach of accessing 2D gapless phases by studying their quasi-1D descendants on ladder geometries, relying heavily on large-scale DMRG calculations. In fact, we have already established[22–24] that for two, three and four legs, the DBM phase itself is the stable ground state of a boson ring-exchange model analogous to equation (6). Here, we take the important first step of placing the electron ring *t*–*J*–*K* model on the two-leg ladder in search of a two-leg descendant of the *d*-wave metal.

For concreteness, we now consider the model, equation (6), on the two-leg ladder (see Fig. 1) at a generic electron density of $\rho = N_e/(2L_x) = 1/3$, where $N_e = N_{c\uparrow} + N_{c\downarrow}$ is the total number of electrons and L_x is the length of our two-leg ladder (that is, the system has $L_x \times 2$ total sites). At this density $\rho = 1/3 < 1/2$ on the two-leg ladder, the non-interacting ground state is a spin-singlet wherein electrons of each spin partially fill the bonding band ($k_y = 0$), leaving the antibonding band ($k_y = \pi$) empty. Thus, for $t \gg K$, we expect the system to be in a simple one-band metallic state, which is a two-leg analogue of the Fermi liquid. Formally speaking, this phase is a conventional Luttinger liquid with two 1D gapless modes (central charge $c = 2$). For moderate values of ring exchange, $K \gtrsim t$, we anticipate the unconventional non-Fermi-liquid *d*-wave metal to be a candidate ground state. On the two-leg ladder at this density, the *d*-wave metal phase has characteristic band-filling configurations for the d_1, d_2 and $f_{\uparrow/\downarrow}$ partons as shown in Fig. 2: d_1 partially fills both bonding and antibonding bands, whereas d_2 and $f_{\uparrow/\downarrow}$ fill only the bonding band. (The d_1 and d_2 configurations constitute the phase denoted "DBL[2,1]" in ref. 22.) In a mean-field approximation in which the partons do not interact, the system has five 1D gapless modes corresponding to the five total partially filled bands. However, in the strong-coupling limit of the full quasi-1D gauge theory (see the Supplementary Information for details), two orthonormal linear combinations of the original five modes are rendered massive, leaving an unconventional Luttinger liquid with $c = 3$ gapless modes.

We now provide extensive numerical evidence that this two-leg descendant of the *d*-wave metal exists as the ground state of the *t*–*J*–*K* model over a wide region of the phase diagram. We summarize these results in Fig. 3 by presenting the full phase diagram in the parameters *K*/*t* versus *J*/*t* as obtained by DMRG calculations on length $L_x = 24$ and 48 systems at electron density $\rho = 1/3$. For small *K*, we find a conventional one-band (spinful) Luttinger liquid phase which is a two-leg analogue of the Fermi-liquid metal. For moderate *J* and upon increasing *K*, the system goes into the unconventional non-Fermi-liquid *d*-wave metal phase, which is the main focus of this work. The phase boundaries in Fig. 3, all of which represent strong first-order transitions, were determined by measuring several standard momentum-space correlation functions in the DMRG (see the Supplementary Information for details): the electron momentum distribution function $\left\langle c_{\boldsymbol{q}s}^\dagger c_{\boldsymbol{q}s} \right\rangle$, the density–density structure factor $\left\langle \delta n_{\boldsymbol{q}} \delta n_{-\boldsymbol{q}} \right\rangle$, and the spin–spin structure factor $\left\langle \boldsymbol{S}_{\boldsymbol{q}} \cdot \boldsymbol{S}_{-\boldsymbol{q}} \right\rangle$.

For concreteness, we now focus on the cut along $J/t = 2$ in Fig. 3 for a 48×2 system with $N_e = 32$ electrons. We take one point deep within the conventional one-band metal at $K/t = 0.5$ and the other

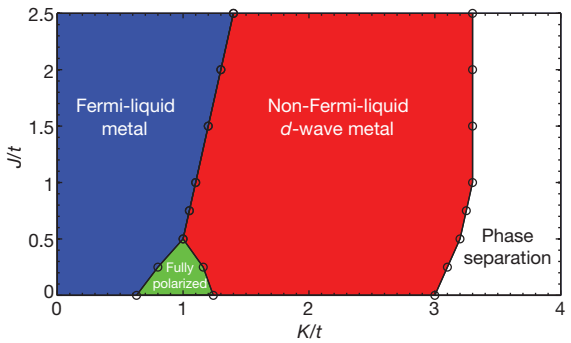


Figure 3 | Phase diagram of the *t*–*J*–*K* electron ring-exchange model at electron density $\rho = 1/3$ on the two-leg ladder. In addition to the conventional one-band metal ('Fermi-liquid metal') and exotic 'non-Fermi-liquid *d*-wave metal', there are two other realized phases. For small *J*, there is an intermediate phase with fully polarized electrons. For large *K*, owing to the inherently attractive nature of ring-exchange interactions[22], the system generally phase separates along the ladder.

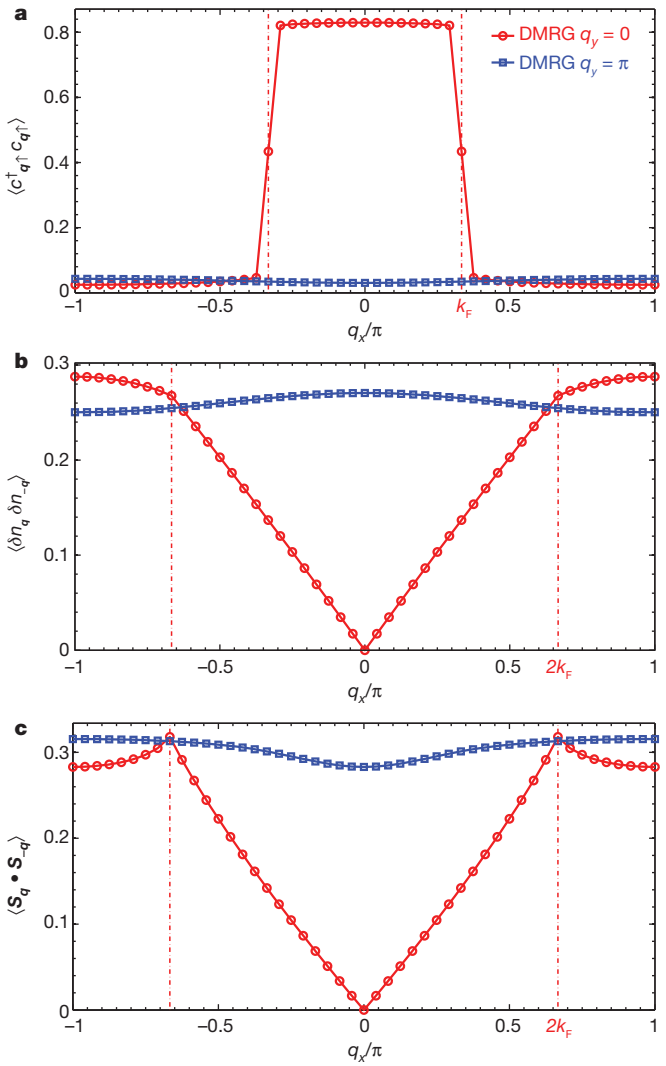


Figure 4 | DMRG measurements in the conventional Luttinger liquid phase at $J/t = 2$ and $K/t = 0.5$. We show the electron momentum distribution function (**a**), the density–density structure factor (**b**) and the spin–spin structure factor (**c**). The important wavevectors k_F and $2k_F$, as described in the text, are highlighted by vertical dashed-dotted lines.

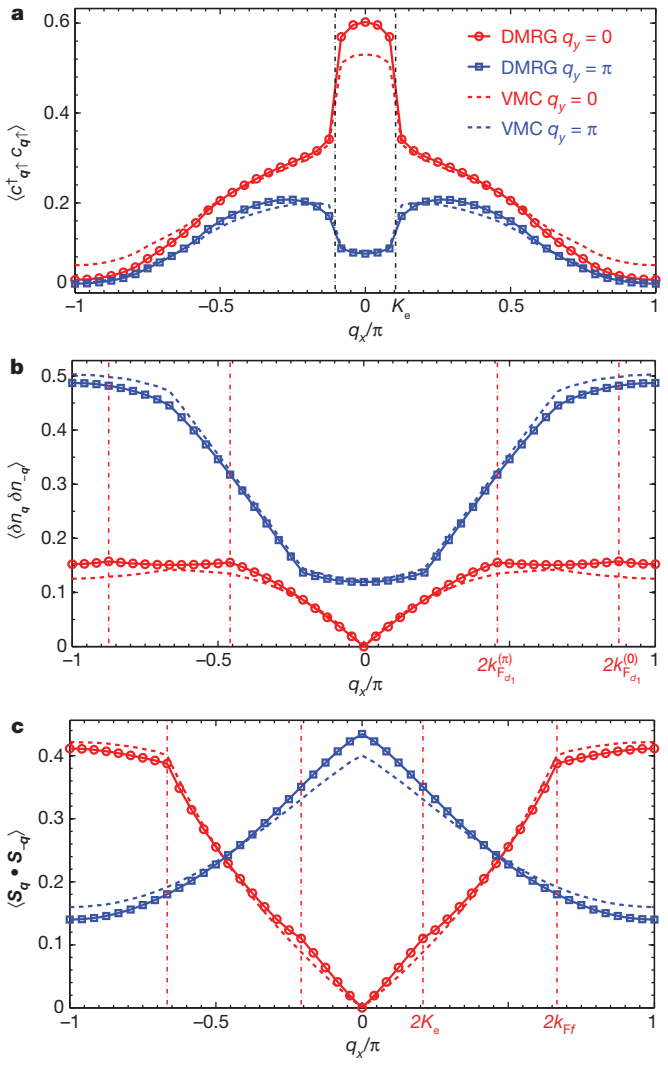


Figure 5 | DMRG measurements in the unconventional *d*-wave metal phase at $J/t = 2$ and $K/t = 1.8$. We show the same quantities as in Fig. 4. Here, we also show the matching VMC measurements using a *d*-wave metal trial wavefunction, depicted in Fig. 2.

point deep within the exotic *d*-wave metal at $K/t = 1.8$. First focusing on the former case, in Fig. 4 we show DMRG measurements characteristic of the conventional Luttinger liquid. The ground state is a spin-singlet with a sharp singularity in the electron momentum distribution function at $q_y = 0$ and $q_x = k_F = \pi N_{c\uparrow}/L_x = 8 \times 2\pi/48$, which is a usual Fermi wavevector determined solely from the electron density. The density–density and spin–spin structure factors at $q_y = 0$ also exhibit familiar features at $q_x = 0$ and $q_x = 2k_F = 16 \times 2\pi/48$, both characteristic of an ordinary one-band metallic state with gapless charge and spin modes[27]. We stress that, even with the constraint of no double-occupancy and non-zero $K/t = 0.5$ and $J/t = 2$, the interacting electronic system is still qualitatively very similar to the two-leg free Fermi gas; analogously, the 2D Fermi liquid is in many ways qualitatively similar to the 2D free Fermi gas. In both cases, the main differences are basically quantitative and are well understood[1,27].

We turn now to the characteristic point within the *d*-wave metal phase at $J/t = 2$ and $K/t = 1.8$. In Fig. 5, we show a set of DMRG measurements at this point, as well as measurements corresponding to a variational wavefunction chosen such that its singular features best reproduce the DMRG data (see the Supplementary Information for details of our VMC methods). The selected *d*-wave metal wavefunction is depicted schematically in Fig. 2. Specifically, we have the following parton Fermi wavevectors: $2k^{(0)}_{Fd_1} = 21 \times 2\pi/48$, $2k^{(\pi)}_{Fd_1} = 11 \times 2\pi/48$, $2k_{Fd_2} = 32 \times 2\pi/48$, and $2k_{Ff} = 16 \times 2\pi/48$. The overall agreement between the DMRG and VMC measurements is very compelling, and we now summarize our understanding of these results from the perspective of *d*-wave metal theory.

In sharp contrast to the conventional Luttinger liquid, the electron momentum distribution function now has singularities for both $q_y = 0$ and $q_y = \pi$ at a wavevector $q_x = K_e \equiv [k^{(0)}_{Fd_1} - k^{(\pi)}_{Fd_1}]/2$. This wavevector corresponds to a composite electron made from a combination of parton fields consisting of a right-moving d_1 parton, a left-moving d_2 parton, and a right-moving spinon: $d^{(q_y)}_{1R} d_{2L} f_{\uparrow R}$. In fact, these 'enhanced electrons' can be guessed from simple 'Amperian rules'[3,43,44] in our quasi-1D gauge theory, as described in detail in the Supplementary Information.

The corresponding density–density and spin–spin structure factors, displayed in Fig. 5b and c, also show nontrivial behaviour. We expect the density–density structure factor to be sensitive to each parton configuration individually and thus have singular features at various '$2k_F$' parton wavevectors (see refs 21, 23 and the Supplementary Information). In the DMRG measurements, the most noticeable features are at $q_y = 0$ and $q_x = 2k^{(0)}_{Fd_1}$, $2k^{(\pi)}_{Fd_1}$, which allow

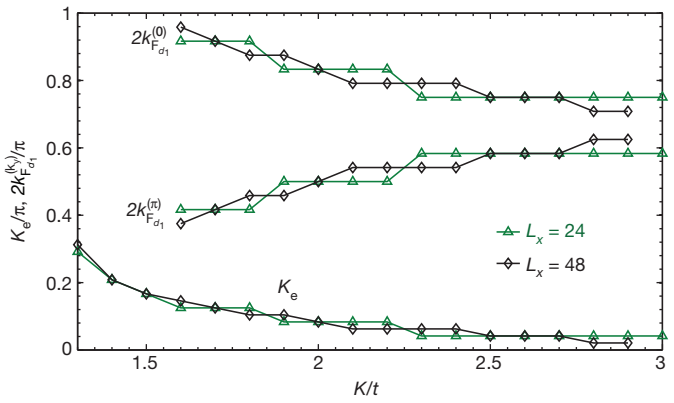


Figure 6 | Evolution of singular wavevectors in the *d*-wave metal phase. At fixed $J/t = 2$ and varying K/t, we show the location of the dominant singular wavevector K_e in the electron momentum distribution function (see Fig. 5a), as well as the wavevectors identified as $2k^{(0)}_{Fd_1}$ and $2k^{(\pi)}_{Fd_1}$ in the density–density structure factor (see Fig. 5b). These calculations were done with DMRG.

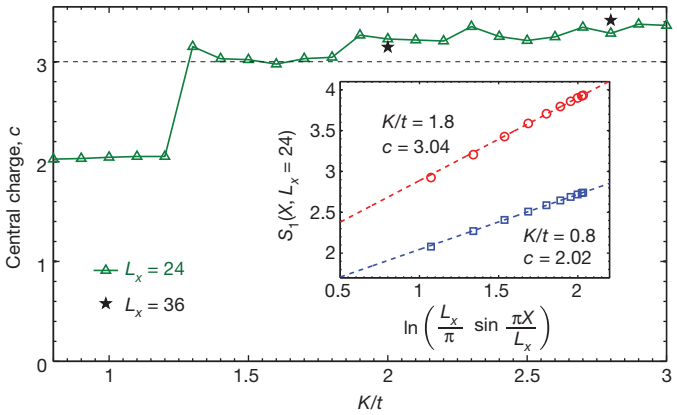


Figure 7 | Central charge *c* as a function of interaction *K*/*t*. By measuring the von Neumann entanglement entropy S_1 in the DMRG, we calculate the effective central charge c at fixed $J/t = 2$ and varying K/t. There is a dramatic jump from $c \simeq 2$ to $c \simeq 3$ at the transition, as predicted by our theory. Data for two example points—$K/t = 0.8$ and 1.8—are shown in the inset, where X is the number of rungs in each bipartition. (See also the Supplementary Information.)

us to read off directly the realized d_1 parton configuration (see Fig. 2). The lack of these features in the VMC data, as well as the lack of analogous features at $q_x = 2k_{Fd_2}$ in the DMRG data, can be understood within our gauge theory framework as presented in the Supplementary Information, where we also note that our wavefunction is only a caricature of the full theory. Finally, the spin–spin structure factor at $q_y = 0$ not only has a familiar, expected feature at $q_x = 2k_{Ff}$ coming from the spinon, but also remarkably contains a feature at $q_x = 2K_e$ that can be thought of as a '$2k_F$' wavevector from the dominant 'electron' in Fig. 5a. All in all, as we further explain in the Supplementary Information, the DMRG measurements are consistent, even on a fine quantitative level, with being in a stable non-Fermi-liquid *d*-wave metal phase.

We note that the wavevector K_e depends on the interaction strength K/t because the wavevectors $k^{(0)}_{Fd_1}$ and $k^{(\pi)}_{Fd_1}$ vary with ring exchange[22]. In Fig. 6, we show at $J/t = 2$ evolution with K/t of the wavevector K_e, that is, the location of the sharp steps in the electron momentum distribution function (see Fig. 5a), as determined by DMRG. Given that the momentum-space 'volume' enclosed by these singular features depends on the interaction K/t and is not simply determined by the total density of electrons, we may confidently say that the *d*-wave metal violates Luttinger's volume theorem[7]. In fact, the very notion of a single 'Fermi surface' is actually ambiguous in the *d*-wave metal phase. We also show in Fig. 6, for those values of K/t at which they are discernible, the wavevectors $2k^{(0)}_{Fd_1}$ and $2k^{(\pi)}_{Fd_1}$ as identified by features in the DMRG-measured density–density structure factor at $q_y = 0$ (see Fig. 5b). For all points, the locations of the identified features satisfy the nontrivial identity $K_e = [2k^{(0)}_{Fd_1} - 2k^{(\pi)}_{Fd_1}]/4$, as predicted by our theory.

A remarkable property of the *d*-wave metal state found in the DMRG is that it has prominent critical *d*-wave Cooper pairs residing on the diagonals, as anticipated earlier from the ring energetics (see the Supplementary Information). Such Cooper pair correlations have the slowest power-law decay of all the discussed observables, including the electron Green's function. This is in stark contrast with a conventional metal and suggests that the *d*-wave metal phase has some incipient *d*-wave superconductivity in two dimensions.

As a final piece of evidence that the realized DMRG phase is in fact the *d*-wave metal, we have measured the number of 1D gapless modes, that is, the effective central charge c, via scaling of the bipartite entanglement entropy[45,46] in the DMRG and VMC[47,48] wavefunctions. As explained above, we expect $c = 2$ in the conventional Luttinger liquid and $c = 3$ in the *d*-wave metal. See the Supplementary Information for a detailed comparison of the DMRG and VMC entropy measurements, where we show that the DMRG–VMC agreement is just as impressive as it is for the more traditional measurements of Fig. 5. The effective central charge versus K/t at $J/t = 2$ as determined by the DMRG is shown in Fig. 7. Indeed, these measurements indicate that $c \simeq 2$ in the conventional one-band metal, whereas $c \simeq 3$ in the exotic *d*-wave metal. Because $c = 3 > 2$, our putative *d*-wave metal phase clearly cannot be understood as an instability out of the conventional one-band metal, but also, because $c = 3 < 4$, the critical bonding and antibonding electrons in Fig. 5a cannot be reproduced by any perturbative treatment starting from free electrons[31] (see also the Supplementary Information).

Discussion and outlook

Here we have presented strong evidence for the stability of a two-leg descendant of our exotic strange-metal phase, the *d*-wave metal, and we conclude with an outlook on exciting future work. Firstly, our present two-leg *d*-wave metal treatment is readily extendable to systems with more legs. Ref. 24 established the stability of the *d*-wave Bose metal, the main ingredient of the *d*-wave metal, on three- and four-leg ladders. Thus, we do not envision any conceptual obstacles in the way of realizing a similar result for the *d*-wave metal. However, we do anticipate that adding more legs will be very challenging numerically for the DMRG owing to the large amount of spatial entanglement present in the *d*-wave metal and Fermi liquid—this is also the current limitation preventing modern 2D tensor network state methods from attacking such problems[49].

With the goal of connecting to experiments, it would be desirable to perform a detailed energetics study of the *t*–*J*–*K* model in two dimensions and explore the applicability of such models to strongly correlated materials. By studying 2D variational wavefunctions based on the *d*-wave metal, it should be possible to compare physical properties with experimentally observed strange metals, such as that in the copper oxides. This could include various instabilities of the *d*-wave metal, such as spinon pairing as a model of a pseudogap metal or chargon pairing as a model of an "orthogonal metal", discussed recently[50]. The *d*-wave sign structure already inherent in the non-superconducting parent *d*-wave metal suggests that there may be incipient *d*-wave superconductivity of the copper-oxide variety, which is particularly exciting. Although we have stressed its Luttinger volume violation as a characteristic non-Fermi liquid property of the *d*-wave metal, we note that the 2D phase will also have no Landau quasiparticle as well as exhibit non-Fermi-liquid-like thermodynamics and transport. Comparing these predictions with properties of real strange metals would be interesting. In the end, however, we stress the conceptual nature of the present study, and hope that our ideas may open up new avenues for thinking about non-Fermi liquid electronic fluids.

Received 1 August; accepted 29 October 2012.
Published online 19 December 2012.

1. Baym, G. & Pethick, C. *Landau Fermi-Liquid Theory: Concepts and Applications* (Wiley-VCH, Germany, 1991).
2. Schofield, A. J. Non-Fermi liquids. *Contemp. Phys.* **40,** 95–115 (1999).
3. Lee, P. A., Nagaosa, N. & Wen, X.-G. Doping a Mott insulator: physics of high-temperature superconductivity. *Rev. Mod. Phys.* **78,** 17–85 (2006).
4. Boebinger, G. S. An abnormal normal state. *Science* **323,** 590–591 (2009).
5. Stewart, G. R. Non-Fermi-liquid behavior in *d*- and *f*-electron metals. *Rev. Mod. Phys.* **73,** 797–855 (2001).
6. Gegenwart, P., Si, Q. & Steglich, F. Quantum criticality in heavy-fermion metals. *Nature Phys.* **4,** 186–197 (2008).
7. Luttinger, J. M. Fermi surface and some simple equilibrium properties of a system of interacting fermions. *Phys. Rev.* **119,** 1153–1163 (1960).
8. Anderson, P. W. & Zou, Z. "Normal" tunneling and "normal" transport: diagnostics for the resonating-valence-bond state. *Phys. Rev. Lett.* **60,** 132–135 (1988).
9. Varma, C. M., Littlewood, P. B., Schmitt-Rink, S., Abrahams, E. & Ruckenstein, A. E. Phenomenology of the normal state of Cu-O high-temperature superconductors. *Phys. Rev. Lett.* **63,** 1996–1999 (1989).
10. Senthil, T. Critical Fermi surfaces and non-Fermi liquid metals. *Phys. Rev. B* **78,** 035103 (2008).
11. Faulkner, T., Iqbal, N., Liu, H., McGreevy, J. & Vegh, D. Strange metal transport realized by gauge/gravity duality. *Science* **329,** 1043–1047 (2010).
12. Sachdev, S. Holographic metals and the fractionalized Fermi liquid. *Phys. Rev. Lett.* **105,** 151602 (2010).
13. Anderson, P. W. The resonating valence bond state in La_2CuO_4 and superconductivity. *Science* **235,** 1196–1198 (1987).
14. Baskaran, G., Zou, Z. & Anderson, P. W. The resonating valence bond state and high-Tc superconductivity – a mean field theory. *Solid State Commun.* **63,** 973–976 (1987).
15. Nagaosa, N. & Lee, P. A. Normal-state properties of the uniform resonating-valence-bond state. *Phys. Rev. Lett.* **64,** 2450–2453 (1990).
16. Lee, P. A. & Nagaosa, N. Gauge theory of the normal state of high-T_c superconductors. *Phys. Rev. B* **46,** 5621–5639 (1992).
17. Wen, X.-G. & Lee, P. A. Theory of underdoped cuprates. *Phys. Rev. Lett.* **76,** 503–506 (1996).
18. Anderson, P. W., Baskaran, G., Zou, Z. & Hsu, T. Resonating valence-bond theory of phase transitions and superconductivity in La_2CuO_4-based compounds. *Phys. Rev. Lett.* **58,** 2790–2793 (1987).
19. Feigelman, M. V., Geshkenbein, V. B., Ioffe, L. B. & Larkin, A. I. Two-dimensional Bose liquid with strong gauge-field interaction. *Phys. Rev. B* **48,** 16641–16661 (1993).
20. Cooper, R. A. *et al.* Anomalous criticality in the electrical resistivity of $La_{2-x}Sr_xCuO_4$. *Science* **323,** 603–607 (2009).
21. Motrunich, O. I. & Fisher, M. P. A. *d*-wave correlated critical Bose liquids in two dimensions. *Phys. Rev. B* **75,** 235116 (2007).
22. Sheng, D. N., Motrunich, O. I., Trebst, S., Gull, E. & Fisher, M. P. A. Strong-coupling phases of frustrated bosons on a two-leg ladder with ring exchange. *Phys. Rev. B* **78,** 054520 (2008).
23. Block, M. S. *et al.* Exotic gapless Mott insulators of bosons on multileg ladders. *Phys. Rev. Lett.* **106,** 046402 (2011).
24. Mishmash, R. V. *et al.* Bose metals and insulators on multileg ladders with ring exchange. *Phys. Rev. B* **84,** 245127 (2011).
25. White, S. R. Density matrix formulation for quantum renormalization groups. *Phys. Rev. Lett.* **69,** 2863–2866 (1992).
26. White, S. R. Density-matrix algorithms for quantum renormalization groups. *Phys. Rev. B* **48,** 10345–10356 (1993).
27. Giamarchi, T. *Quantum Physics in One Dimension* (Oxford Univ. Press, 2003).
28. Shankar, R. Bosonization: how to make it work for you in condensed matter. *Acta Phys. Polon. B* **26,** 1835–1867 (1995).
29. Lin, H.-H., Balents, L. & Fisher, M. P. A. Exact SO(8) symmetry in the weakly-interacting two-leg ladder. *Phys. Rev. B* **58,** 1794–1825 (1998).
30. Fjærestad, J. O. & Marston, J. B. Staggered orbital currents in the half-filled two-leg ladder. *Phys. Rev. B* **65,** 125106 (2002).
31. Balents, L. & Fisher, M. P. A. Weak-coupling phase diagram of the two-chain hubbard model. *Phys. Rev. B* **53,** 12133–12141 (1996).
32. Ceperley, D., Chester, G. V. & Kalos, M. H. Monte Carlo simulation of a many-fermion study. *Phys. Rev. B* **16,** 3081–3099 (1977).
33. Gros, C. Physics of projected wavefunctions. *Ann. Phys.* **189,** 53–88 (1989).
34. Hellberg, C. S. & Mele, E. J. Phase diagram of the one-dimensional *t-J* model from variational theory. *Phys. Rev. Lett.* **67,** 2080–2083 (1991).
35. Ceperley, D. M. Fermion nodes. *J. Stat. Phys.* **63,** 1237–1267 (1991).
36. Halperin, B. I., Lee, P. A. & Read, N. Theory of the half-filled Landau level. *Phys. Rev. B* **47,** 7312–7343 (1993).
37. Laughlin, R. B. Anomalous quantum Hall effect: an incompressible quantum fluid with fractionally charged excitations. *Phys. Rev. Lett.* **50,** 1395–1398 (1983).
38. Normand, B. & Oleś, A. M. Circulating-current states and ring-exchange interactions in cuprates. *Phys. Rev. B* **70,** 134407 (2004).
39. Sheng, D. N., Motrunich, O. I. & Fisher, M. P. A. Spin Bose-metal phase in a spin-1/2 model with ring exchange on a two-leg triangular strip. *Phys. Rev. B* **79,** 205112 (2009).
40. Block, M. S., Sheng, D. N., Motrunich, O. I. & Fisher, M. P. A. Spin Bose-metal and valence bond solid phases in a spin-1/2 model with ring exchanges on a four-leg triangular ladder. *Phys. Rev. Lett.* **106,** 157202 (2011).
41. Imada, M. & Miyake, T. Electronic structure calculation by first principles for strongly correlated electron systems. *J. Phys. Soc. Jpn* **79,** 112001 (2010).
42. Troyer, M. & Wiese, U.-J. Computational complexity and fundamental limitations to fermionic quantum Monte Carlo simulations. *Phys. Rev. Lett.* **94,** 170201 (2005).
43. Polchinski, J. Low-energy dynamics of the spinon-gauge system. *Nucl. Phys. B* **422,** 617–633 (1994).
44. Altshuler, B. L., Ioffe, L. B. & Millis, A. J. Low-energy properties of fermions with singular interactions. *Phys. Rev. B* **50,** 14048–14064 (1994).
45. Calabrese, P. & Cardy, J. Entanglement entropy and quantum field theory. *J. Stat. Mech.* **2004,** P06002 (2004).
46. Calabrese, P., Campostrini, M., Essler, F. & Nienhuis, B. Parity effects in the scaling of block entanglement in gapless spin chains. *Phys. Rev. Lett.* **104,** 095701 (2010).
47. Hastings, M. B., González, I., Kallin, A. B. & Melko, R. G. Measuring Renyi entanglement entropy in quantum Monte Carlo simulations. *Phys. Rev. Lett.* **104,** 157201 (2010).
48. Zhang, Y., Grover, T. & Vishwanath, A. Entanglement entropy of critical spin liquids. *Phys. Rev. Lett.* **107,** 067202 (2011).
49. Corboz, P., Orús, R., Bauer, B. & Vidal, G. Simulation of strongly correlated fermions in two spatial dimensions with fermionic projected entangled-pair states. *Phys. Rev. B* **81,** 165104 (2010).
50. Nandkishore, R., Metlitski, M. A. & Senthil, T. Orthogonal metals: the simplest non-Fermi liquids. *Phys. Rev. B* **86,** 045128 (2012).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank T. Senthil, R. Kaul, L. Balents, S. Sachdev, A. Vishwanath and P. Lee for discussions. This work was supported by the NSF under the KITP grant PHY05-51164 and the MRSEC programme under award number DMR-1121053 (H.-C.J.), the NSF under grants DMR-1101912 (M.S.B., R.V.M., J.R.G. and M.P.A.F.), DMR-1056536 (M.S.B.), DMR-0906816 and DMR-1205734 (D.N.S.), DMR-0907145 (O.I.M.), and by the Caltech Institute of Quantum Information and Matter, an NSF Physics Frontiers Center with the support of the Gordon and Betty Moore Foundation (O.I.M. and M.P.A.F.). We also acknowledge support from the Center for Scientific Computing from the CNSI, MRL: an NSF MRSEC award (DMR-1121053), and an NSF grant (CNS-0960316).

Author Contributions All authors made significant contributions to the research underlying this paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to R.V.M. (mishmash@physics.ucsb.edu).

ARTICLE

doi:10.1038/nature11711

Genomic variation landscape of the human gut microbiome

Siegfried Schloissnig[1]*, Manimozhiyan Arumugam[1]*, Shinichi Sunagawa[1]*, Makedonka Mitreva[2], Julien Tap[1], Ana Zhu[1], Alison Waller[1], Daniel R. Mende[1], Jens Roat Kultima[1], John Martin[2], Karthik Kota[2], Shamil R. Sunyaev[3], George M. Weinstock[2] & Peer Bork[1,4]

Whereas large-scale efforts have rapidly advanced the understanding and practical impact of human genomic variation, the practical impact of variation is largely unexplored in the human microbiome. We therefore developed a framework for metagenomic variation analysis and applied it to 252 faecal metagenomes of 207 individuals from Europe and North America. Using 7.4 billion reads aligned to 101 reference species, we detected 10.3 million single nucleotide polymorphisms (SNPs), 107,991 short insertions/deletions, and 1,051 structural variants. The average ratio of non-synonymous to synonymous polymorphism rates of 0.11 was more variable between gut microbial species than across human hosts. Subjects sampled at varying time intervals exhibited individuality and temporal stability of SNP variation patterns, despite considerable composition changes of their gut microbiota. This indicates that individual-specific strains are not easily replaced and that an individual might have a unique metagenomic genotype, which may be exploitable for personalized diet or drug intake.

With the increasing availability of individual human genomes, various theoretical and practical aspects of genomic variation can be deduced for individuals and the human population as a whole[1,2]. Like sequenced human genomes, the number of human gut metagenomes (currently mostly derived from Illumina shotgun sequencing of stool samples) is increasing exponentially. Given the importance of the gut microbiota in human health[3,4] and a growing number of studies reporting associations between gut microbiota and diseases[5–8], an understanding of genomic variation in gut microbial populations will probably trigger applications towards human well-being and disease.

For example, in the common gut commensal bacterium *Escherichia coli*, just three point mutations in two genes can confer clinically relevant antibiotic resistance[5], and natural variation in a single gene can lead to pathogenic adaptation[8]. Even within pathogenic species in the gut, closely related coexisting strains can exhibit different pathogenic potentials due to minor genomic variation[7]. These examples illustrate how genomic variation within gut microbes could confer phenotypes that require personalized care or treatment of the host.

Studies based on 16S ribosomal RNA gene surveys or whole metagenome shotgun sequencing characterized taxonomic and functional compositions of healthy individuals' and intestinal bowel disease patients' gut microbiota at the genus or species level[6,9–12]. Variation in taxonomic abundance as well as functions encoded by these gut microbiota have been described between individuals[6,11] and used to stratify individuals according to their gut community compositions into enterotypes[13]. However, genomic variation within species, which leads to their phenotypic diversity and adaptations to different environments, has only been studied in a few taxa, such as *Citrobacter* spp[7].

An early landmark study on a small data set described metagenomic variation in an acidic biofilm microbiome of low complexity[14]. The population structure for one species in that habitat was studied and positive selection was observed in some genes[15]. Another recent study resolved multiple clinical isolates of methicillin-resistant *Staphylococcus aureus* and delineated its epidemiology and microevolution based on genomic variation[16]. With the availability of hundreds of deeply sequenced human gut metagenomes[9,11,17], sufficient data are becoming available for quantitative analyses of the genetic structure of complex microbial communities, allowing the study of many species at the same time.

Here we analysed 1.56 terabases of sequence data from 252 stool samples from 207 individuals (Supplementary Table 1 and Supplementary Notes) obtained from the MetaHIT project (71 Danish, 39 Spanish; all sampled once[11]), the NIH Human Microbiome Project (94 US samples; 51 individuals sampled once, 41 sampled twice and 2 sampled three times[9]), and Washington University (three US samples; all sampled once[12]). Our goals were to: (1) develop a framework for genomic variation analysis using metagenomic shotgun data; (2) gather basic knowledge on the genomic variation landscape in gut metagenomes; and (3) gain insights into the individuality, temporal stability and biogeography of metagenomic variation.

Framework for metagenomic variation analysis

We used 1,497 prokaryotic genomes to generate a set of reference genomes (Supplementary Table 2) for the analysis of genomic variation in gut microbial species in 252 samples (on average 6.2 ± 4.1 gigabases (Gb) were analysed). Pairwise comparisons of 40 universal marker genes[18,19] identified in these genomes were performed to create a set of 929 clusters based on a 95% DNA identity threshold recommended for identifying species[20]. The genome recruiting the highest number of reads in a cluster was selected as reference for that species (see Methods and Supplementary Information).

Using the same 95% identity threshold, we mapped 7.4 billion metagenomic reads (42% of the total, 91% thereof uniquely) with an average length of 80 base pairs (bp) to the 929 reference genomes (Supplementary Tables 1 and 3). To avoid mapping artefacts (for example, caused by high coverage of prophages), we required ≥40%

[1]European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany. [2]The Genome Institute, Washington University School of Medicine, 4444 Forest Park Avenue, St Louis, Missouri 63108, USA. [3]Division of Genetics, Department of Medicine, Brigham & Women's Hospital & Harvard Medical School, Boston, Massachusetts 02115, USA. [4]Max Delbrück Centre for Molecular Medicine, D-13092 Berlin, Germany.

*These authors contributed equally to this work.

of each reference genome to be covered by reads (corresponding to the gene content similarity between two strains of *E. coli*[21]). The resulting 101 prevalent species with base-pair coverage from 12× to 32,400× (Fig. 1 and Supplementary Fig. 1) were subjected to genomic variation analysis.

To enable comparative analyses in multiple metagenomes and to identify low-frequency variants not detected when analysing them individually, we used multi-sample calling[22] to identify SNPs, short insertions/deletions (indels; 1–50 bp) and structural variations (>50 bp) in each genome, although structural variations were largely underestimated due to small insert sizes (Supplementary Information). We only called variants with allele frequency ≥1% (the conventional definition of polymorphism[2]) and supported by ≥4 reads. False-positive rates were estimated at 0.71% for SNPs and 1.04% for structural variations (Supplementary Information, Supplementary Tables 4 and 5, and Supplementary Fig. 2).

Genomic variation in prevalent gut species

We identified 10.3 million SNPs in 101 genomes (3.1% of the total 329 Mb) across 252 samples from 207 subjects, almost as many as the 14.4 million SNPs recently identified in 179 human genomes[2]. Within an individual the rate was lower (on average 1.21%, see Supplementary Table 6), yet SNPs kb^{-1} increased with base-pair coverage when samples were pooled (Supplementary Fig. 3). We also identified 107,991 indels and 1,051 structural variants in these 101 species (Supplementary Information). Their relative ratios to SNPs (10,485 short indels and 102 structural variants per million SNPs) were robust across species and individuals (Supplementary Fig. 4). Subsequent analyses were restricted to SNPs due to their orders of magnitude higher count over other variation types.

We annotated the genes of the prevalent genomes using orthologous groups from eggNOG[23] (Supplementary Information) and found that the orthologous groups with the highest SNP density were enriched in functions related to conjugal transfer of antibiotic resistance (Supplementary Table 7). For example, the orthologous group with the highest average SNP density across samples was the clindamycin resistance transfer factor *btgA* (NOG119724), required for conjugative transmission of plasmids. Mutations commonly accrued from the process of conjugation may account for increased diversity among conjugation-associated functions[24]. Additionally, CRISPR-associated proteins, responsible for conferring resistance in bacteria, were also found among the orthologous groups with high SNP densities (Supplementary Information and Supplementary Table 7).

The large number of SNPs provided the opportunity to compare, for the first time at such scale, the evolution of different coexisting species across a large cohort of individuals. To evaluate selective constraints in these species in their natural habitat, we estimated the ratio of non-synonymous to synonymous polymorphism rates[25,26] (pN/pS) within each species in every sample (Fig. 1 and Supplementary Information). pN/pS characterizes selective constraint at the level of a population contrary to the more commonly used dN/dS that characterizes it between individual species[26]. To validate pN/pS ratios, we estimated genetic variation using the sample-size-independent nucleotide diversity π, and found that π is highly correlated with SNPs kb^{-1} (Fig. 1 and Supplementary Fig. 5). The derived measures of $\pi(N)/\pi(S)$ and π(non-degenerate sites)/π(fourfold degenerate sites), the latter of

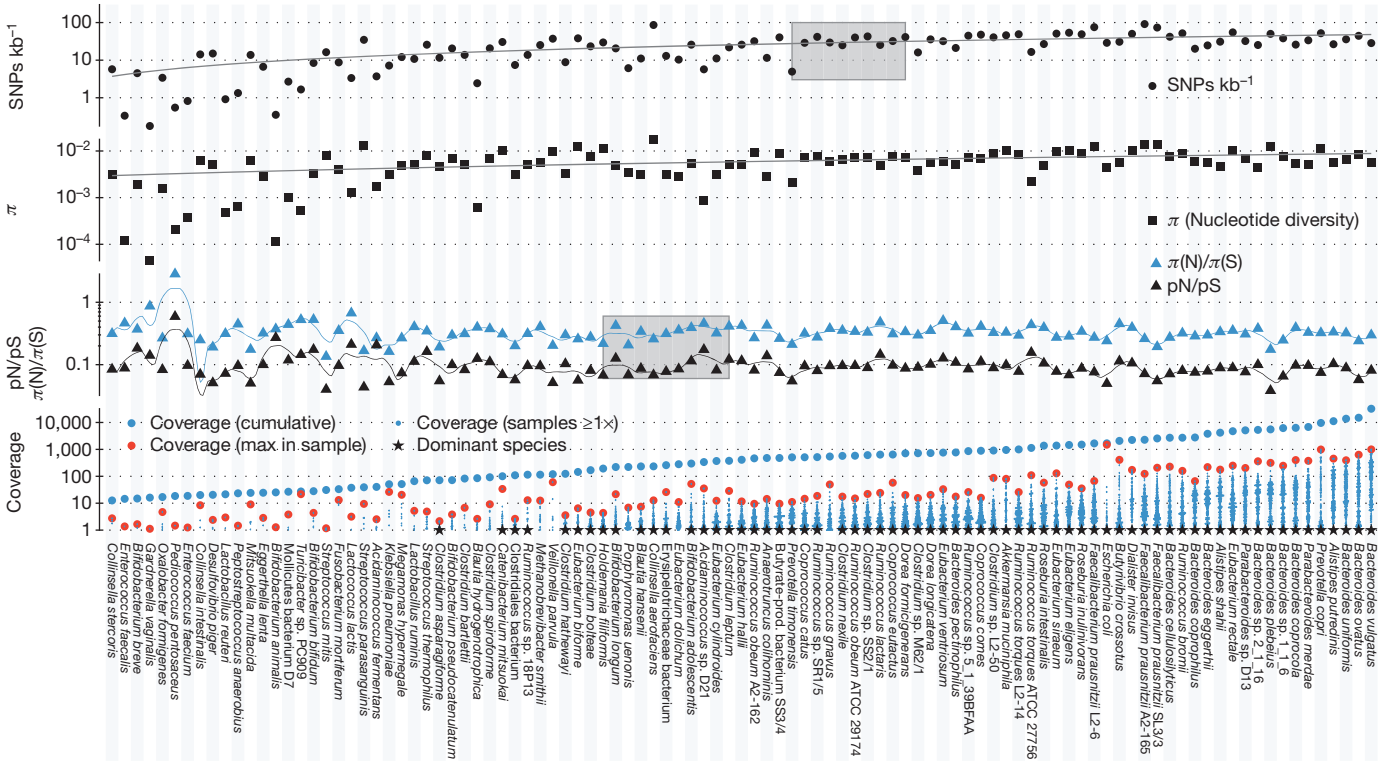


Figure 1 | Genomic variation statistics for 101 gut microbial species prevalent in 252 samples from 207 individuals. Genomic variation statistics were calculated for 101 prevalent gut microbial species, operationally defined as having ≥10× cumulative (over all samples) base-pair coverage with at least one sample exhibiting a genome coverage of ≥40%. The 66 dominant species (indicated by an asterisk), which account for 99% of the mapped reads, were used for analyses that required high base-pair coverage. Species names are given without strain specifications unless this would result in duplicate entries. The blue point cloud plots show the coverages (≥1×) in all samples, with the blue dot above indicating the cumulative coverage and the red dot the maximum coverage across all samples. Grey shaded areas indicate the level of base-pair coverage at which abundance effects have only minor effects on SNPs kb^{-1} and pN/pS ratios of the pooled samples (Supplementary Information). SNP counts appear to saturate at approximately 500×, with minor increases at higher coverages probably due to the sampling of rare variants at low rates. In individual samples, pN/pS is largely stable from a coverage of 10× onward (Supplementary Fig. 7), corresponding to approximately 200× cumulative coverage in our sample set. Nucleotide diversity π follows SNPs kb^{-1} closely, as does the derived measure of $\pi(N)/\pi(S)$ with respect to pN/pS.

which is less dependent on specific properties of mutation spectra such as transition and transversion ratios, were coherent with pN/pS (Supplementary Fig. 6).

The pN/pS ratio of a genome within a sample remained stable at coverages higher than 10× (Supplementary Fig. 7)—yet another indication of few false SNP calls—and was on average 0.11, but varied considerably for different species (0.04 to 0.58) in accordance with dN/dS ratios estimated independently in a number of interspecies comparisons between closely related bacteria and archaea[27,28].

pN/pS across gut species and individuals

Because meaningful comparison of genomic variation requires both breadth (across samples) and depth (in number of base pairs) of sequencing, we focused on the 66 most dominant species that attracted >99% of the reads (Fig. 1). Their relatively low pN/pS ratios were constant across different hosts (Fig. 2a and Supplementary Table 8), which may indicate similar selective constraints across individuals. Thus, the evolution of gut species is probably dominated by long-term purifying selection and drift rather than rapid adaptations to specific host environments. The wide range of these ratios across species may suggest that different gut species face different evolutionary constraints.

To investigate how different gut species respond to the pressure from the gut environment, we compared the pN/pS ratios of individual genes in *Roseburia intestinalis* and *Eubacterium eligens*, which differed considerably in their overall mean pN/pS ratios (0.236 (*R. intestinalis*) versus 0.131 (*E. eligens*) from 106 and 147 samples, respectively) despite having comparable average base-pair coverages (Supplementary Information). Whereas 75% of the genes in *R. intestinalis* had systematically higher pN/pS ratios compared to their orthologues in *E. eligens*, few others showed considerable deviations (Fig. 2b and Supplementary Table 9), indicative of differing evolutionary constraints for these genes. For example, *galK*, the gene encoding galactokinase, an essential enzyme in the Leloir pathway for galactose metabolism in most organisms[29], was among the lowest in terms of pN/pS ratio in *R. intestinalis* but among the highest in *E. eligens* (0.03 and 0.48, respectively; Fig. 2b, c). Although present in *E. eligens*, this gene may not exert its main function (see also ref. 30), as *E. eligens* cannot ferment galactose, nor the galactose-containing disaccharides lactose and melibiose[31]. On the other hand, *R. intestinalis* is known to ferment melibiose[32], indicating that its *galK* gene is functional (Supplementary Information). Thus, the same gene can be under tight negative selection in one species but under more relaxed negative selection in another.

Our framework allowed us additionally to obtain information on all genes in each sample (Supplementary Information). As expected, we found that housekeeping genes had usually lower pN/pS ratios. For example, the DNA-dependent RNA polymerase β-subunit gene was consistently among the genes exhibiting the lowest pN/pS ratios across samples and species (Supplementary Table 10). Less obvious examples included genes related to type IV secretion systems used to transfer DNA between microbes[33] and involved in host interactions of both pathogenic[33] and commensal bacteria[34], specifically in anti-inflammatory responses and immune modulation[35]. The low pN/pS ratio of genes related to type IV secretion systems suggests that maintaining genome plasticity and antibiotic resistance through conjugative transposition is essential in the constantly changing environment of the gut and that the interaction with the host immune system is under purifying selection (Supplementary Table 10). We also found a few conserved unknown, but apparently gut-microbe-specific, proteins that exhibited low pN/pS ratios, suggesting that they perform important yet hitherto unexplored functions (Supplementary Table 11).

Among the genes or orthologous groups with consistently the highest pN/pS ratios were many transposases and antimicrobial resistance genes including the gut-specific gene bile salt hydrolase (BSH)[36] (Supplementary Table 10). Conjugated bile acids (CBAs) secreted by the hosts repress microbial growth and upregulate the host mucosal defence system. BSHs are involved in the initial reaction in the metabolism of CBAs by gut microbes[36]. Their high pN/pS ratio may be indicative of the genomic plasticity necessary to metabolize and survive the variety of different bile acids present in the gut[37].

Temporal stability of individual SNP patterns

Several studies on adult human gut microbial samples from a few individuals have suggested that within-individual differences over time are smaller compared to between-individual differences in microbial species composition and abundance[38–40]. Within a larger cohort, individuality of host-associated microbiota has been reported on the basis of 16S rRNA gene profiling of fingertip-associated communities[41], whereas other studies on a few samples have investigated the persistence of specific strains over time[42,43]. However, intra-species variation at nucleotide resolution at the whole-genome level and accompanying changes in species abundances within the human gut over long time periods (>1 year) have not been studied yet in large cohorts. It is unclear if the concept of resident strains is common to other prevalent species, if host-specific strains are retained over time, and how fast they evolve inside the gut environment.

To explore these questions, we used 88 gut metagenomes from 43 healthy US subjects (a subset of our cohort) from whom at least two samples were obtained at different time points with no antibiotics treatment in between (Supplementary Table 12). To measure how similar the subpopulations (strains) of the dominant species were between two samples, we estimated the fixation indices (F_{ST}) between the populations (Supplementary Information). Because this measure depends on allele frequencies, which cannot be determined accurately at low base-pair coverage, we also estimated the fraction of alleles shared between the samples out of all polymorphic sites (only 49 genomes that accrued 40% genome coverage in at least two samples were used and genomes with >10× base-pair coverage were down-sampled to 10×; Supplementary Information and Supplementary Fig. 3). Because the fraction of shared SNPs depends on the number of variable sites, we developed a heuristic allele sharing similarity score that takes into account both the number of variable sites and the fraction of shared alleles (Supplementary Information).

When we compared all 252 samples, F_{ST} was significantly lower and allele sharing significantly stronger between different samples from the same individuals than between samples from different individuals (Mann–Whitney *U*-test: $P < 0.001$ for both; see Fig. 3a, b and Supplementary Information). The same trend was observed, albeit much weaker, based on species compositions (Fig. 3c), in line with previous observations from microbial composition-based results[38–40]. Intra-individual variation being smaller than inter-individual variation does not require that samples from the same individual are more similar to each other than to any other sample in the tested cohort. Our results showed for both measures of variation similarity that all but one of the 88 multi-time-point samples had the highest similarity to another sample from the same individual, which was not true when comparing species abundance over time (Fig. 3c and Supplementary Information). This indicates that species abundance in gut microbiota cannot serve as a fingerprint of an individual whereas variation patterns might.

We also tested whether differences in F_{ST} and allele sharing decreased over time, which may indicate a divergence of the strains or a horizontal transmission of strains from the environment; however, the individual-specific variation patterns remained stable over all of the time intervals monitored (Fig. 3). Although this stability should be verified for longer periods as well as when antibiotic treatment or other gut microbiota-challenging events have taken place, our observation indicates that healthy human individuals retain specific strains (see also Supplementary Tables 12 and 13) for at least 1 year.

In contrast to the strong evidence for individuality and temporal stability of SNP patterns, we did not observe a significant geographical

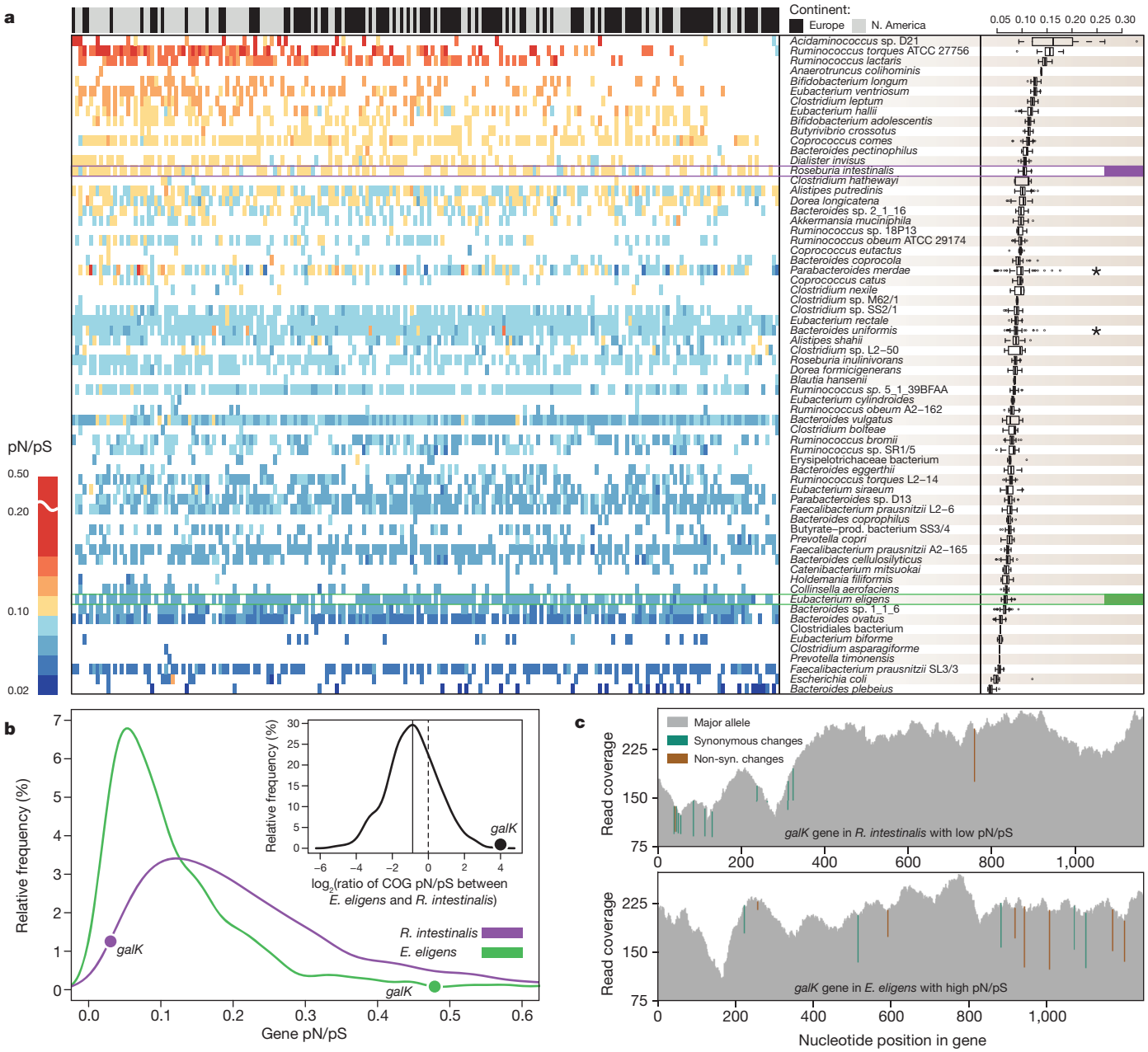


Figure 2 | pN/pS ratios of 66 dominant species reveal more variation between species than between individuals. **a**, A heat map of pN/pS ratios for the 66 dominant species (rows) and 207 individuals (columns; only the first time-point per individual) is shown and summarized by species (box plots on the right). Rows and columns are ordered by their mean pN/pS ratios, which vary considerably between species, but have a tighter bandwidth across samples. Two genomes that are exceptions to this trend (indicated by an asterisk) might indicate higher strain diversity. The panel above the heat map indicates the continent of residence for each individual. A significant difference was found in the mean pN/pS ratios between the two continents, although this is probably an effect of lower sequencing depths of European samples (Supplementary Table 8) that leads to missing data points in some samples (see, for example, top-right corner). **b**, The distributions of average pN/pS ratios of individual genes from *Roseburia intestinalis* and *Eubacterium eligens* (both highlighted in **a**) illustrate that, although base-pair coverages are similar, the pN/pS ratio of *R. intestinalis* is higher in general. The relative pN/pS ratios of orthologous groups in the two species are shown in the inset, with the average $\log_2$ ratio indicated by the solid line and the random expectation by the dashed line. Outliers can be revealed this way, like the galactokinase gene (*galK*), the pN/pS of which is among the lowest in *R. intestinalis* and the highest in *E. eligens*. **c**, Illustration of low and high pN/pS ratios in *galK* genes from *R. intestinalis* (top panel) and *E. eligens* (bottom panel). The cumulative read coverage is shown in grey with synonymous (green) and non-synonymous (brown) changes marked at the nucleotide positions they occur.

separation between European and US samples (Supplementary Fig. 8 and Supplementary Table 14). This indicates that long-term horizontal transmission of at least some dominant gut microbial strains cause geographical mixing over time. The strongest continental separation, based on F_{ST}, was seen in *Bacteroides coprocola* (Fig. 4), which was also the only genome with sufficient amounts of data that showed continental separation based on the allele sharing score (Supplementary Fig. 8 and Supplementary Tables 14 and 15).

Discussion

We have established a framework for gut microbial genomic variation analysis using metagenomic data and identified in a single analysis—involving 252 stool samples from 207 human individuals—almost as many SNPs in the human gut microbiome as the 1000 Genomes Project recorded in 179 human individuals over several years[2]. The stable pN/pS ratios of gut microbial species across individuals suggest that host conditions (such as diet, genetic differences, and immune

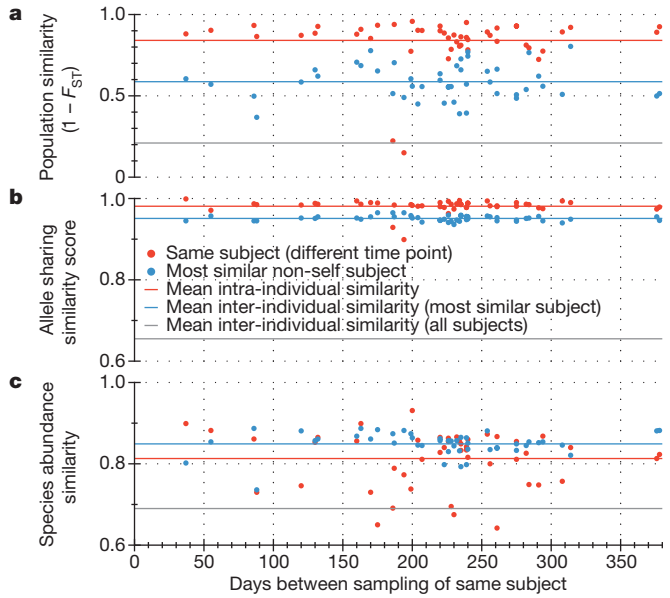


Figure 3 | Individuality and temporal stability of genomic variation patterns. **a**–**c**, Samples from 43 individuals that were sampled at different time intervals (red dots) were compared with the most similar sample from a different individual (blue dots) in terms of population similarity that takes allele frequencies into account (**a**), allele sharing similarity score that takes SNP counts and the ratio of shared SNPs into account (**b**) (Supplementary Information) and species abundance similarity measured using the Jensen–Shannon distance[13] (**c**). The most similar sample is the one with the lowest F_{ST} value in **a**, the highest allele sharing similarity score in **b** and the lowest Jensen–Shannon distance in **c**. The three similarity measures are plotted against the number of days between the sampling time points. The mean across all intra-individual, best inter-individual and all inter-individual similarities is shown as red, blue and grey lines, respectively. For both population similarity and allele sharing similarity between samples from the same individual, all but one sample (resulting in two outliers due to comparisons with two other time-points, see Supplementary Table 12) shared the highest similarity with another sample of the same individual, providing strong evidence for individuality of SNP sharing patterns. No decline of similarity over time could be observed.

tolerance) have a minor influence on the evolution of species compared to constraints common to the human population (such as gut physiology, anaerobic conditions and pH). In the 66 dominant species, the analysis of more than 229,000 genes comprising about 8,000 orthologous groups pinpointed consistently fast or slow evolving genes across individuals (Supplementary Table 10). However, further studies are needed to interpret different selection types at the gene level.

The availability of time-point data revealed that individual-specific variation patterns were remarkably stable over time (Fig. 3a, b), which was much less the case for similarities in species abundance—for almost 60% of the samples, a sample from a different individual was the most similar (Fig. 3c). Thus, the metagenomic variation patterns observed here support the hypothesis that a healthy individual retains specific strains for extended periods of time. This suggests that each individual has a metagenomic variation profile that could be unique even in very large cohorts. It should be noted that the maximal sampling period was only 1 year, and 43 individuals might not be sufficient to trace horizontal transmissions of strains. The likelihood of the latter is supported by the apparent absence of clear continental stratification (despite different sampling and sequencing protocols of the European and the US samples), although for one out of eight species analysed, *Bacteroides coprocola*, we provide preliminary evidence (Fig. 4 and Supplementary Fig. 8). Geographical stratification has been described for *Helicobacter pylori*[44,45], and weak but detectable signals have also been observed in some bacterial pathogens[46,47]. Thus, we expect more gut microbial stratification patterns to emerge when larger data sets under standardized sampling and sequencing protocols become available, although it remains to be tested which factors (such as geographical separation, diseases, host-genetic and life-style/diet factors) shape the distribution of gut microbial strains and segregating SNPs within the population. The absence of clear geographical stratification implies that stable differences in variation patterns of gut species are not explained by large-scale structures of local microbial populations. They may rather be a result of genetic drift due to population bottlenecks that could occur not only during the colonization of the infant gut but also by processes causing community shifts during adult life stages, followed by a rapid population growth accompanied by purifying selection. This model suggests that the source of genetic variation in human gut microbial populations is less likely to be new mutations within the host than the variation in the initial colonizing populations or transmissions from the environment. This would imply that most allelic variants analysed in this study segregate at timescales greatly exceeding human generation time.

The introduction of large-scale variation analysis in metagenomic data of complex communities and the discovery of individual metagenomic variation profiles open up several applications. It is now possible to screen *in silico* for many pathogenic or antibiotics resistance variants in the population. Once a sample has been analysed, the data can also be used in the future given the temporal stability of SNP profiles. As it took years to identify marker genes and variations for diseases or phenotypes in the human genome, the variation landscape uncovered here can only be seen as the beginning to find molecular biomarkers including particular variants that reveal useful information for human health and well-being.

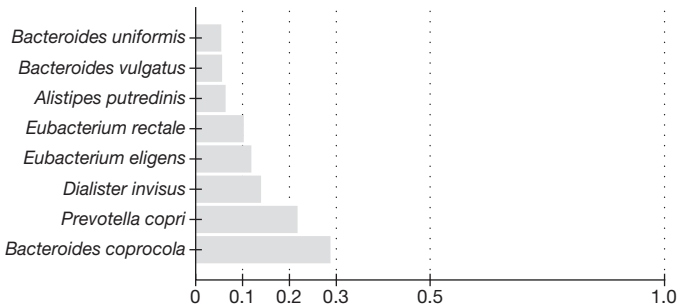


Figure 4 | Inter-continental comparison of gut microbial species. Between-continent F_{ST} values for eight genomes with ≥10 samples representing each continent are shown. *Bacteroides coprocola* was the species with the highest F_{ST} value, implying a separation between the *B. coprocola* populations in Europe and North America (see also Supplementary Information; all data are available in Supplementary Table 14).

METHODS SUMMARY

Mapping to non-redundant genomes. A reference genome set representing 929 species was derived from a total of 1,497 published prokaryotic genomes, based on a median sequence identity of 95% in 40 universal, single-copy marker genes[18,19]. Metagenomic reads from 252 samples were aligned to these 929 genomes using the same 95% sequence identity cutoff.

Coverage. For each genome, we calculated the sample-specific base-pair coverage and the number of bases of the genome covered by at least one read. For a genome to be considered we required a cumulative depth of coverage of ≥10× across all samples. To remove species that are not present in our cohort, yet attract reads due to highly conserved regions, we required at least 40% breadth of the genome coverage (the criterion for the species to be considered present) from at least one sample.

Variation detection. We performed SNP calling on the pooled samples and only considered bases with a quality score ≥15. We required SNPs to be supported by ≥4 reads and to occur with a frequency of ≥1%. Structural variations were detected using Pindel[48]. False-positive rates in SNP and structural variation detection were estimated using nonsense and frameshift mutations in 40 essential single copy marker genes.

π and F_{ST}. We estimated nucleotide diversity (π) and fixation indices (F_{ST}) on the basis of allele frequencies.

pN/pS ratio. SNPs occurring in coding regions were classified as synonymous or non-synonymous. Genes from the non-redundant genomes were annotated using eggNOG orthologous groups allowing calculation of pN/pS ratios at the level of genomes, orthologous groups and genes.

Pairwise sample comparisons. Similarity in strain populations between two samples was estimated using a similarity score based on shared SNPs and using F_{ST}.

Received 25 January; accepted 25 October 2012.
Published online 5 December 2012.

1. International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449,** 851–861 (2007).
2. The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467,** 1061–1073 (2010).
3. Backhed, F., Ley, R. E., Sonnenburg, J. L., Peterson, D. A. & Gordon, J. I. Host-bacterial mutualism in the human intestine. *Science* **307,** 1915–1920 (2005).
4. Hooper, L. V., Midtvedt, T. & Gordon, J. I. How host-microbial interactions shape the nutrient environment of the mammalian intestine. *Annu. Rev. Nutr.* **22,** 283–307 (2002).
5. Bagel, S., Hüllen, V., Wiedemann, B. & Heisig, P. Impact of *gyrA* and *parC* mutations on quinolone resistance, doubling time, and supercoiling degree of *Escherichia coli*. *Antimicrob. Agents Chemother.* **43,** 868–875 (1999).
6. Eckburg, P. B. *et al.* Diversity of the human intestinal microbial flora. *Science* **308,** 1635–1638 (2005).
7. Morowitz, M. J. *et al.* Strain-resolved community genomic analysis of gut microbial colonization in a premature infant. *Proc. Natl Acad. Sci. USA* **108,** 1128–1133 (2011).
8. Sokurenko, E. V. *et al.* Pathogenic adaptation of *Escherichia coli* by natural variation of the FimH adhesin. *Proc. Natl Acad. Sci. USA* **95,** 8922–8926 (1998).
9. The Human Microbiome Project Consortium. A framework for human microbiome research. *Nature* **486,** 215–221 (2012).
10. Lay, C. *et al.* Colonic microbiota signatures across five northern European countries. *Appl. Environ. Microbiol.* **71,** 4153–4155 (2005).
11. Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464,** 59–65 (2010).
12. Turnbaugh, P. J. *et al.* A core gut microbiome in obese and lean twins. *Nature* **457,** 480–484 (2009).
13. Arumugam, M. *et al.* Enterotypes of the human gut microbiome. *Nature* **473,** 174–180 (2011).
14. Tyson, G. W. *et al.* Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428,** 37–43 (2004).
15. Allen, E. E. *et al.* Genome dynamics in a natural archaeal population. *Proc. Natl Acad. Sci. USA* **104,** 1883–1888 (2007).
16. Harris, S. R. *et al.* Evolution of MRSA during hospital transmission and intercontinental spread. *Science* **327,** 469–474 (2010).
17. Peterson, J. *et al.* The NIH Human Microbiome Project. *Genome Res.* **19,** 2317–2323 (2009).
18. Ciccarelli, F. D. *et al.* Toward automatic reconstruction of a highly resolved tree of life. *Science* **311,** 1283–1287 (2006).
19. Sorek, R. *et al.* Genome-wide experimental determination of barriers to horizontal gene transfer. *Science* **318,** 1449–1452 (2007).
20. Konstantinidis, K. T. & Tiedje, J. M. Prokaryotic taxonomy and phylogeny in the genomic era: advancements and challenges ahead. *Curr. Opin. Microbiol.* **10,** 504–509 (2007).
21. Touchon, M. *et al.* Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. *PLoS Genet.* **5,** e1000344 (2009).
22. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genet.* **43,** 491–498 (2011).
23. Muller, J. *et al.* eggNOG v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations. *Nucleic Acids Res.* **38,** D190–D195 (2010).
24. Kunz, B. A. & Glickman, B. W. The infidelity of conjugal DNA transfer in *Escherichia coli*. *Genetics* **105,** 489–500 (1983).
25. Simmons, S. L. *et al.* Population genomic analysis of strain variation in *Leptospirillum* group II bacteria involved in acid mine drainage formation. *PLoS Biol.* **6,** e177 (2008).
26. McDonald, J. H. & Kreitman, M. Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* **351,** 652–654 (1991).
27. Friedman, R., Drake, J. W. & Hughes, A. L. Genome-wide patterns of nucleotide substitution reveal stringent functional constraints on the protein sequences of thermophiles. *Genetics* **167,** 1507–1512 (2004).
28. Novichkov, P. S., Wolf, Y. I., Dubchak, I. & Koonin, E. V. Trends in prokaryotic evolution revealed by comparison of closely related bacterial and archaeal genomes. *J. Bacteriol.* **191,** 65–73 (2009).
29. Frey, P. A. The Leloir pathway: a mechanistic imperative for three enzymes to change the stereochemical configuration of a single carbon in galactose. *FASEB J.* **10,** 461–470 (1996).
30. Kuhner, S. *et al.* Proteome organization in a genome-reduced bacterium. *Science* **326,** 1235–1240 (2009).
31. Holdeman, L. V. & Moore, W. E. C. New genus, *Coprococcus*, twelve new species, and emended descriptions of four previously described species of bacteria from human feces. *Int. J. Syst. Bacteriol.* **24,** 260–277 (1974).
32. Duncan, S. H., Hold, G. L., Barcenilla, A., Stewart, C. S. & Flint, H. J. *Roseburia intestinalis* sp. nov., a novel saccharolytic, butyrate-producing bacterium from human faeces. *Int. J. Syst. Evol. Microbiol.* **52,** 1615–1620 (2002).
33. Alvarez-Martinez, C. E. & Christie, P. J. Biological diversity of prokaryotic type IV secretion systems. *Microbiol. Mol. Biol. Rev.* **73,** 775–808 (2009).
34. Nagai, H. & Roy, C. R. Show me the substrates: modulation of host cell function by type IV secretion systems. *Cell Microbiol.* **5,** 373–383 (2003).
35. Kelly, D., Conway, S. & Aminov, R. Commensal gut bacteria: mechanisms of immune modulation. *Trends Immunol.* **26,** 326–333 (2005).
36. Jones, B. V., Begley, M., Hill, C., Gahan, C. G. M. & Marchesi, J. R. Functional and comparative metagenomic analysis of bile salt hydrolase activity in the human gut microbiome. *Proc. Natl Acad. Sci. USA* **105,** 13580–13585 (2008).
37. Begley, M., Hill, C. & Gahan, C. G. M. Bile salt hydrolase activity in probiotics. *Appl. Environ. Microbiol.* **72,** 1729–1738 (2006).
38. Caporaso, J. G. *et al.* Moving pictures of the human microbiome. *Genome Biol.* **12,** R50 (2011).
39. Dethlefsen, L. & Relman, D. A. Incomplete recovery and individualized responses of the human distal gut microbiota to repeated antibiotic perturbation. *Proc. Natl Acad. Sci. USA* **108,** 4554–4561 (2011).
40. Zoetendal, E. G., Akkermans, A. D. & De Vos, W. M. Temperature gradient gel electrophoresis analysis of 16S rRNA from human fecal samples reveals stable and host-specific communities of active bacteria. *Appl. Environ. Microbiol.* **64,** 3854–3859 (1998).
41. Fierer, N. *et al.* Forensic identification using skin bacterial communities. *Proc. Natl Acad. Sci. USA* **107,** 6477–6481 (2010).
42. Tenaillon, O., Skurnik, D., Picard, B. & Denamur, E. The population genetics of commensal *Escherichia coli*. *Nature Rev. Microbiol.* **8,** 207–217 (2010).
43. Jernberg, C., Lofmark, S., Edlund, C. & Jansson, J. K. Long-term ecological impacts of antibiotic administration on the human intestinal microbiota. *ISME J.* **1,** 56–66 (2007).
44. Suzuki, R., Shiota, S. & Yamaoka, Y. Molecular epidemiology, population genetics, and pathogenic role of *Helicobacter pylori*. *Infect. Genet. Evol.* **12,** 203–213 (2012).
45. Yamaoka, Y. *Helicobacter pylori* typing as a tool for tracking human migration. *Clin. Microbiol. Infect.* **15,** 829–834 (2009).
46. Achtman, M. & Wagner, M. Microbial diversity and the genetic nature of microbial species. *Nature Rev. Microbiol.* **6,** 431–440 (2008).
47. Morelli, G. *et al.* *Yersinia pestis* genome sequencing identifies patterns of global phylogenetic diversity. *Nature Genet.* **42,** 1140–1143 (2010).
48. Ye, K., Schulz, M. H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25,** 2865–2871 (2009).

Supplementary Information is available in the online version of the paper.

Acknowledgements The authors are grateful to J. Korbel and the members of the Bork group at EMBL for discussions and assistance, especially S. Powell for performing some of the computations. We thank the EMBL IT core facility and Y. Yuan for managing the high-performance computing resources. We would like to thank J. I. Gordon for providing three of the samples used. We are also grateful to the European MetaHIT consortium and the NIH Common Fund Human Microbiome Project Consortium for generating and making available the data sets used in this study. The research leading to these results has received funding from EMBL, the European Community's Seventh Framework Programme via the MetaHIT (HEALTH-F4-2007-201052) and IHMS (HEALTH-F4-2010-261376) grants as well as from the National Institutes of Health grants U54HG003079 and U54HG004968.

Author Contributions P.B. and G.M.W. conceived the study. P.B., M.A., G.M.W. and S.R.S. designed the analyses. Si.S., Sh.S., M.A., M.M., J.T., A.Z., A.W., D.R.M., J.R.K., J.M. and K.K. performed the analyses. M.A., Sh.S., Si.S. and P.B. wrote the manuscript. All authors read and approved the manuscript.

Author Information Single nucleotide polymorphism data have been submitted to dbSNP under accession numbers ss539238913–ss549853572. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to P.B. (bork@embl.de) or G.M.W. (gweinsto@genome.wustl.edu).

ARTICLE

doi:10.1038/nature11724

CCR5 is a receptor for *Staphylococcus aureus* leukotoxin ED

Francis Alonzo III[1], Lina Kozhaya[1]*, Stephen A. Rawlings[1]*, Tamara Reyes-Robles[1]*, Ashley L. DuMont[1], David G. Myszka[4], Nathaniel R. Landau[1], Derya Unutmaz[1,2,3] & Victor J. Torres[1]

Pore-forming toxins are critical virulence factors for many bacterial pathogens and are central to *Staphylococcus aureus*-mediated killing of host cells. *S. aureus* encodes pore-forming bi-component leukotoxins that are toxic towards neutrophils, but also specifically target other immune cells. Despite decades since the first description of staphylococcal leukocidal activity, the host factors responsible for the selectivity of leukotoxins towards different immune cells remain unknown. Here we identify the human immunodeficiency virus (HIV) co-receptor CCR5 as a cellular determinant required for cytotoxic targeting of subsets of myeloid cells and T lymphocytes by the *S. aureus* leukotoxin ED (LukED). We further demonstrate that LukED-dependent cell killing is blocked by CCR5 receptor antagonists, including the HIV drug maraviroc. Remarkably, CCR5-deficient mice are largely resistant to lethal *S. aureus* infection, highlighting the importance of CCR5 targeting in *S. aureus* pathogenesis. Thus, depletion of CCR5$^+$ leukocytes by LukED suggests a new immune evasion mechanism of *S. aureus* that can be therapeutically targeted.

S. aureus is a bacterial pathogen that causes significant morbidity and mortality worldwide. The organism is responsible for a myriad of diseases, from skin and soft-tissue infections, to more invasive diseases including necrotizing pneumonia and sepsis. *S. aureus* secretes several protein products that allow the organism to subvert the host immune system. Such factors include super-antigens, antibody binding proteins, cytolytic peptides and pore-forming cytotoxins[1].

Pore-forming toxins are secreted by a substantial number of pathogenic bacteria[2]. The toxins are secreted as water-soluble monomers that recognize host cell membranes, oligomerize, and insert α-helical or β-barrel pores into the lipid bilayer[2]. Pore formation disrupts osmotic balance and membrane potential, ultimately leading to cell death[2]. *S. aureus* strains that infect humans produce up to four different β-barrel, bi-component, pore-forming toxins (HlgACB, LukED, LukSF-PV/PVL and LukAB/HG) that exhibit a unique tropism for host immune cells and contribute to the greater virulence of *S. aureus*[1,3,4]. The precise repertoire of immune cells targeted by the pore-forming leukotoxins remains to be fully determined. Even now, more than a century since the first description of staphylococcal leukocidal activity[5,6], our understanding of leukotoxin function *in vivo* is limited because of an absence of known host-derived specificity determinants.

CCR5 is required for LukED cytotoxicity

To identify potential leukotoxin receptors, we purified recombinant LukED, LukAB and LukSF-PV and assessed their ability to kill a set of human cell lines[4,7]. Granulocyte-like human cells (PMN-HL60) were killed in 1 h by LukAB and LukSF-PV, but not LukED (Fig. 1a). In contrast, LukED was cytotoxic to a human T-cell line ectopically expressing CCR5 (HUT-R5); whereas another T-cell line (Jurkat), which lacks detectable CCR5, was insensitive (Fig. 1a). This suggested that CCR5 was involved in LukED cytotoxicity towards HUT-R5 cells. Accordingly, when CCR5 amounts were reduced in HUT-R5 cells using lentiviral *CCR5* short hairpin RNA (shRNA), the cells were protected from LukED-mediated killing (Fig. 1b and Supplementary Fig. 1a, b).

Complementary to these findings, ectopic expression of *CCR5* was sufficient to render Jurkat and H9 cells (Supplementary Fig. 1c) susceptible to LukED cytotoxicity (Fig. 1c). As expected, on the basis of the mode of action of the bi-component leukotoxins, CCR5-dependent LukED-mediated cytotoxicity required both LukE and LukD subunits (Supplementary Fig. 2a, b). A human osteosarcoma cell line engineered to constitutively express CCR5 (GHOST.R5 cells)[8] was also sensitive to LukED, but not to LukAB or LukSF-PV (Fig. 1d). The sensitivity of GHOST cells to LukED was specific to CCR5 expression, as overexpression of additional T-cell-specific chemokine receptors (CCR1, CCR2, CCR3, CXCR4, CCR8 and CXCR6) in these cells did not confer susceptibility to LukED (Supplementary Fig. 2c).

CCR5 antagonists block LukED cell killing

CCR5 is a co-receptor required for HIV infection[9–11] and has been targeted with small molecule antagonists aimed at restricting HIV entry into host cells[11]. We found that one such clinically approved receptor antagonist, maraviroc, potently blocked LukED killing of CCR5$^+$ cells (Fig. 1e and Supplementary Fig. 3a) at concentrations similar to those required to block HIV infection (Supplementary Fig. 3b). Similar inhibitory effects were observed with the CCR5 antagonists vicriviroc and TAK-779, as well as chemokines that are natural ligands of CCR5 (Supplementary Fig. 3a, c)[12,13]. We found that maraviroc resulted in complete blockade of LukED pore formation, an essential process for cytotoxicity (Fig. 1f and Supplementary Fig. 3d).

We next investigated whether *S. aureus* was able to kill CCR5$^+$ cells in a LukED-dependent manner. The expression of *lukED* in *S. aureus* is inherently low during *in vitro* growth[7]. However, deletion of the transcription factor Rot, a potent repressor, results in the enhanced expression and production of LukED by *S. aureus*[7]. Thus, to assess *S. aureus* cytotoxicity towards CCR5$^+$ cells, Jurkat or Jurkat-R5 cells were infected with *S. aureus* Δ*rot* (*Sa* LukED$^+$) and *S. aureus*

[1]Department of Microbiology, New York University School of Medicine, New York, New York 10016, USA. [2]Department of Pathology, New York University School of Medicine, New York, New York 10016, USA. [3]Department of Medicine, New York University School of Medicine, New York, New York 10016, USA. [4]Biosensor Tools LLC, Salt Lake City, Utah 84103, USA.
*These authors contributed equally to this work.

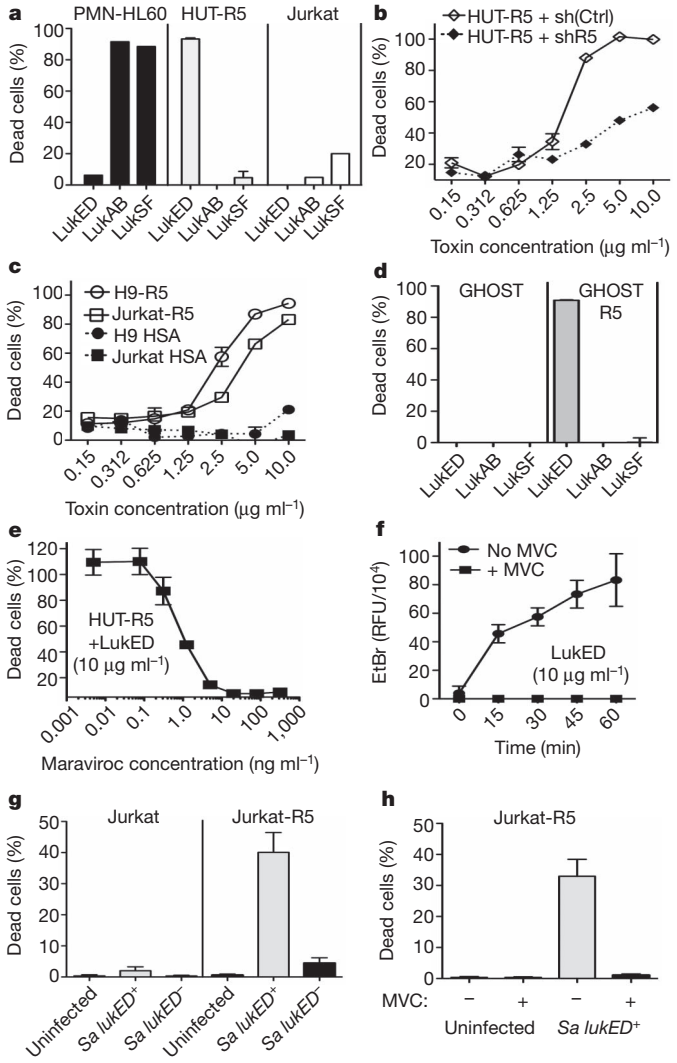


Figure 1 | LukED requires CCR5 for cell killing. **a**, Viability of cells exposed to different leukotoxins (10 μg ml^{-1}). **b**, Viability of HUT-R5 cells transduced with control (Ctrl) or *CCR5* shRNAs. **c**, Viability of Jurkat and H9 cells transduced with CCR5 (-R5) or mouse CD24 (-HSA) followed by treatment with LukED. **d**, Viability of GHOST cells overexpressing CCR5 and treated with indicated leukotoxins. **e**, Viability of HUT-R5 pre-incubated with maraviroc and treated with LukED. **f**, Pore formation, as measured by ethidium bromide uptake, on Jurkat-R5 with or without maraviroc (MVC; 100 ng ml^{-1}) followed by incubation with LukED. **g**, **h**, Viability of Jurkat or Jurkat-R5 cells infected with *S. aureus* (**g**), in the presence or absence of MVC (**h**). Means ± s.d. ($n = 3$) are shown.

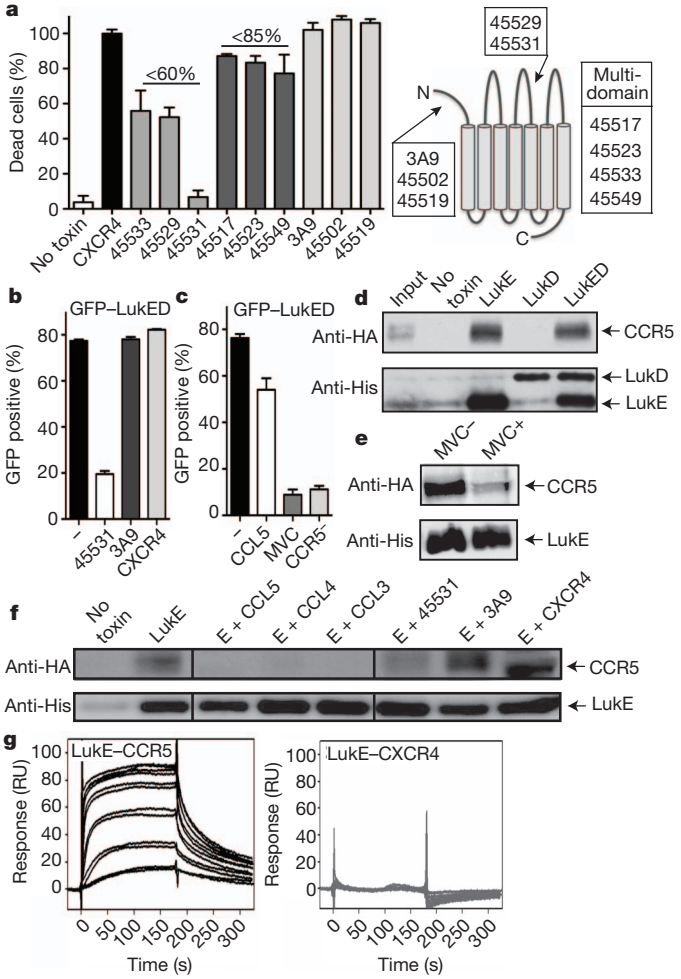


Figure 2 | LukE directly interacts with CCR5. **a**, Viability of cells treated with anti-CCR5 monoclonal antibodies (35 μg ml^{-1}) followed by exposure to LukED (10 μg ml^{-1}). **b**, Membrane association of GFP–LukED (10 μg ml^{-1}) to the surface of primary CD4$^+$CCR5$^+$ T cells with or without the indicated monoclonal antibodies (25 μg ml^{-1}) as determined by fluorescence-activated cell sorting (FACS). **c**, Membrane association of GFP–LukED (10 μg ml^{-1}) on the surface of primary CD4$^+$CCR5$^+$ T cells with or without maraviroc (MVC) (100 ng ml^{-1}), CCL5 (5 μg ml^{-1}) or on CD4$^+$CCR5$^-$ T cells. **d**–**f**, Interaction between His- LukE, LukD, or LukED and HA-CCR5 (**d**), with or without MVC (5 μg ml^{-1}) (**e**), CCL5, CCL4, CCL3 (10 μg ml^{-1}) (**f**) and monoclonal antibodies 45531, 3A9, CXCR4 (35 μg ml^{-1}) (**f**). In **f**, E stands for the LukE toxin subunit. Immunoblots are representative of at least two independent experiments. **g**, Interaction of LukE with CCR5 and CXCR4 by surface plasmon resonance. Representative sensorgrams (**g**) of two experiments performed in duplicate are shown. Where relevant, means ± s.d. ($n = 3$) are shown.

Δ*rot*Δ*lukED* (*Sa* LukED$^-$) mutants. Jurkat-R5 cells were killed by *S. aureus* in a LukED-dependent manner, whereas Jurkats lacking CCR5 were resistant to killing (Fig. 1g). Additionally, Jurkat-R5 killing by *S. aureus* was completely blocked by maraviroc (Fig. 1h).

LukE interacts directly with CCR5

To characterize more precisely the LukED–CCR5 interaction on target cells, we first determined whether monoclonal antibodies specific towards extracellular regions of CCR5 (ref. 14) were sufficient to block toxin activity (Fig. 2a). Antibodies against extracellular loop 2 (ECL-2), but not the amino (N) terminus of the receptor or CXCR4, significantly blocked toxin killing (Fig. 2a) and prevented association of functional green fluorescent protein (GFP)-labelled toxin (Supplementary Fig. 4) with the cell surface of sorted primary human CD4$^+$CCR5$^+$ T cells (Fig. 2b). Furthermore, toxin association with the cell surface of CCR5$^+$ cells was also reduced in the presence of CCL5, and was completely blocked upon addition of maraviroc, similar to CD4$^+$CCR5$^-$ T cells (Fig. 2c). To determine whether LukED interacts with CCR5, pull-down assays were conducted with purified toxin and solubilized CCR5. We found that CCR5 interacted with LukE but not LukD (Fig. 2d). This interaction was significantly reduced in the presence of maraviroc, natural ligands of CCR5, as well as monoclonal antibody 45531 directed against ECL-2, but not 3A9 directed against the N terminus of CCR5 (Fig. 2e, f). Additionally, incubation of LukE (75-fold molar excess) with CCR5$^+$ cells largely blunted native ligand-induced CCR5 signalling as measured by calcium mobilization (Supplementary Fig. 5). LukE itself does not seem to induce CCR5 signalling (Supplementary Fig. 6a, b). Surface plasmon resonance studies with immobilized native CCR5 (ref. 15) and purified LukE or LukD subunits confirmed the pull-down studies and determined that LukE, but not LukD, binds to CCR5 in a

time-dependent and saturable manner, with an apparent dissociation constant (K_d) of 39.6 ± 0.4 nM (Fig. 2g and Supplementary Fig. 7a, b). This interaction was specific, as evidenced by an inability of LukE to bind native CXCR4 (Fig. 2g).

LukED kills CCR5$^+$ myeloid cells and T cells

We next sought to determine the subsets of primary human lymphoid and myeloid cells targeted by LukED. Treatment of blood lymphocytes with LukED resulted in specific depletion of CCR5$^+$ T cells, most of which were effector memory T lymphocytes (Fig. 3a and Supplementary Fig. 8). As with cell lines, the CCR5-dependent killing of primary cells was completely blocked by maraviroc (Fig. 3a and Supplementary Fig. 8). A proportion of individuals of Northern European heritage harbour a 32 base-pair deletion in the *CCR5* gene (Δ32 CCR5), resulting in a truncated protein that cannot be surface localized, thus rendering the CD4$^+$ T cells refractory to HIV infection[11,16,17]. Similarly, primary T cells expanded from a Δ32 CCR5 donor were also resistant to LukED cytotoxicity (Fig. 3b). In keeping with the notion that CCR5 is required for HIV-1 entry into CD4$^+$ T cells[9–11], selective depletion of CCR5$^+$ T cells by LukED suppressed HIV-1 spread (Supplementary Fig. 9).

Memory T cells can be classified into functional subsets on the basis of differential chemokine receptor profiles and cytokine production. Among T-cell subsets, the CCR6$^+$CCR5$^+$ subset produces more interleukin (IL)-17 and interferon (IFN)-γ than CCR6$^+$CCR5$^-$ T cells[18]. Consistent with this association, depletion of CCR5$^+$CD4$^+$ T cells with LukED greatly reduced the proportion of IFN-γ- and IL-17-producing cells compared with purified CD4$^+$ T-cell controls (Fig. 3c, day 0). Incubation with the γc-cytokines IL-7 and IL-15 significantly enhances the proportion of IL-17$^+$ and IL-17$^+$/IFN-γ$^+$ by CCR6$^+$ memory T cells[19]. We found that when human CD4$^+$ T cells were first treated with LukED, followed by 7 days of culture with IL-7 and IL-15, there was a substantial reduction in the induction of IFN-γ and IL-17/IL-22-secreting CCR6$^+$ T cells (Fig. 3c). This finding correlates well with depletion of the CCR6$^+$CCR5$^+$ memory progenitor subset (Supplementary Fig. 10). In addition to Th1 and Th17 effector cells, LukED also killed macrophages and dendritic cells in a CCR5-dependent manner (Fig. 3d).

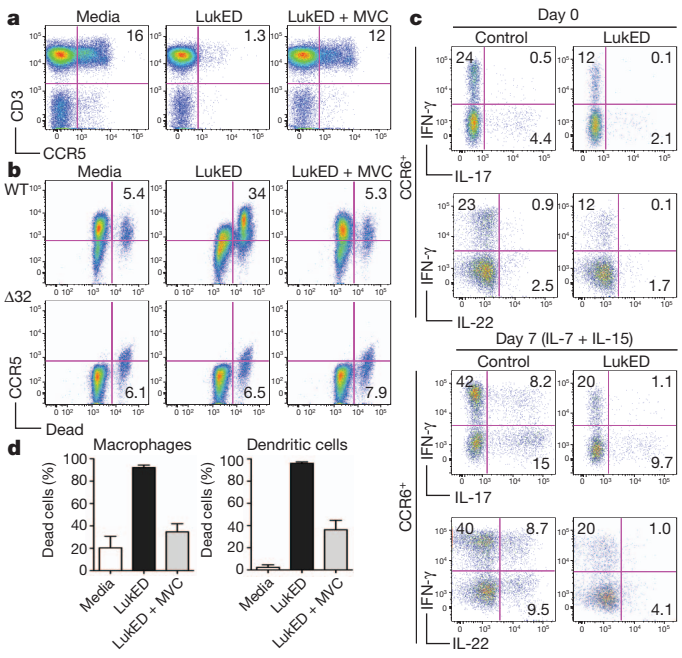


Figure 3 | LukED kills CCR5$^+$ human memory T cells, macrophages and dendritic cells. **a**, Total CCR5$^+$ primary human T cells (CD3$^+$/CCR5$^+$) incubated with media, LukED (2.5 μg ml^{-1}) or maraviroc (MVC; 100 ng ml^{-1}) followed by LukED treatment. **b**, Susceptibility of T cells isolated from a Δ32-CCR5 or WT-CCR5 donor. Cell viability and CCR5 expression evaluated by flow cytometry as in **a**. **c**, Cytokine production of CD4$^+$ T cells with or without LukED treatment (5 μg ml^{-1}) that were stimulated on day 0 with PMA and ionomycin (P + I; top panel) or cultured in media supplemented with IL-7/IL-15 (20 ng ml^{-1}) for 7 days followed by stimulation with P+I (bottom panel). **d**, Viability of monocyte-derived macrophages and dendritic cells incubated with LukED (3.0 μg ml^{-1} with or without MVC). For FACS plots (**a**–**c**), a representative from one of three independent donors is shown. Bar graphs, mean ± s.d. of results from three independent donors.

LukED targets CCR5$^+$ cells *in vivo*

Next we examined the contribution of CCR5 to *S. aureus* pathogenesis and determined the influence of LukED on the targeted killing of CCR5$^+$ cells *in vivo*. We found that murine CCR5 (mCCR5) renders transfected 293T cells fully susceptible to the toxin (Supplementary Fig. 11a, b). Additionally, primary murine macrophages treated with high concentrations of maraviroc were partly protected from toxin-mediated killing, confirming that LukED is directly targeting mCCR5 (Supplementary Fig. 11c). Because maraviroc is potent towards human CCR5 but not mCCR5 (Fig. 1e and Supplementary Fig. 11)[20], we chose to study wild-type (WT) and CCR5-deficient mice with the hypothesis that the latter would be resistant to LukED cytotoxicity. *S. aureus*-elicited lymphocytes and macrophages from WT mice were highly susceptible to purified LukED, whereas lymphocytes and macrophages isolated from *CCR5*$^{-/-}$ mice were markedly resistant (Fig. 4a, b). To validate further that *S. aureus* kills CCR5$^+$ leukocytes *in vivo*, we implemented a peritonitis model in which WT and *CCR5*$^{-/-}$ mice were infected with *S. aureus*. CCR5 surface expression was not required for the initial influx of immune cells to the infection site, as the cells recovered and their profiles were identical among all mice (Supplementary Fig. 12). However, lymphocytes and macrophages elicited *in vivo* in WT mice were more susceptible to *S. aureus* killing than those from the *CCR5*$^{-/-}$ mice (Fig. 4c, d). LukED is associated with *S. aureus* pathogenesis in a murine model of systemic infection[7]. Using this model, *CCR5*$^{-/-}$ mice infected with WT *S. aureus* exhibited significantly reduced bacterial burden in the kidneys than those of infected WT mice (Fig. 4e), a phenotype similar to that observed for mice infected with a *S. aureus* Δ*lukED* mutant[7]. After 96 h, infected *CCR5*$^{-/-}$ mice also exhibited significantly reduced serum pro-inflammatory cytokines and chemokines and showed a commensurate reduction in innate immune cells in the kidney compared with WT mice (Fig. 4f, g), signs consistent with infection resolution. Additionally, when WT mice were challenged systemically with WT or a Δ*lukED* mutant, we observed LukED-dependent killing of CCR5$^+$ macrophages in infected kidneys, consistent with our hypothesis that LukED is capable of targeting CCR5$^+$ leukocytes during infection (Fig. 4h). In support of the importance of CCR5 targeting *in vivo*, the mortality associated with *S. aureus* bloodstream infection was reduced for CCR5-deficient mice, a phenotype similar to that of mice challenged with strains of *S. aureus* lacking *lukED* (Fig. 4i).

Discussion

To our knowledge, CCR5 is the first described cellular receptor that is necessary and sufficient for the killing of mammalian cells by a staphylococcal bi-component leukotoxin. Thus, in addition to HIV, *Toxoplasma gondii* and poxviruses (vaccinia and myxoma)[9,21–24], *S. aureus* can also exploit CCR5 to target immune cells. Interestingly, the Δ32 allele of *CCR5* is thought to have been acquired through selective pressure imparted by a deadly pathogen[25,26]. *Yersinia pestis* or variola virus were postulated as potential driving forces behind this selection, but these hypotheses have either been discounted or remain uncertain in favour of an older selection event incited by an immune-cell-targeting pathogen[24,27]. Our findings put forth the possibility that resistance to *S. aureus* leukotoxins may have influenced the selection of the Δ32 allele.

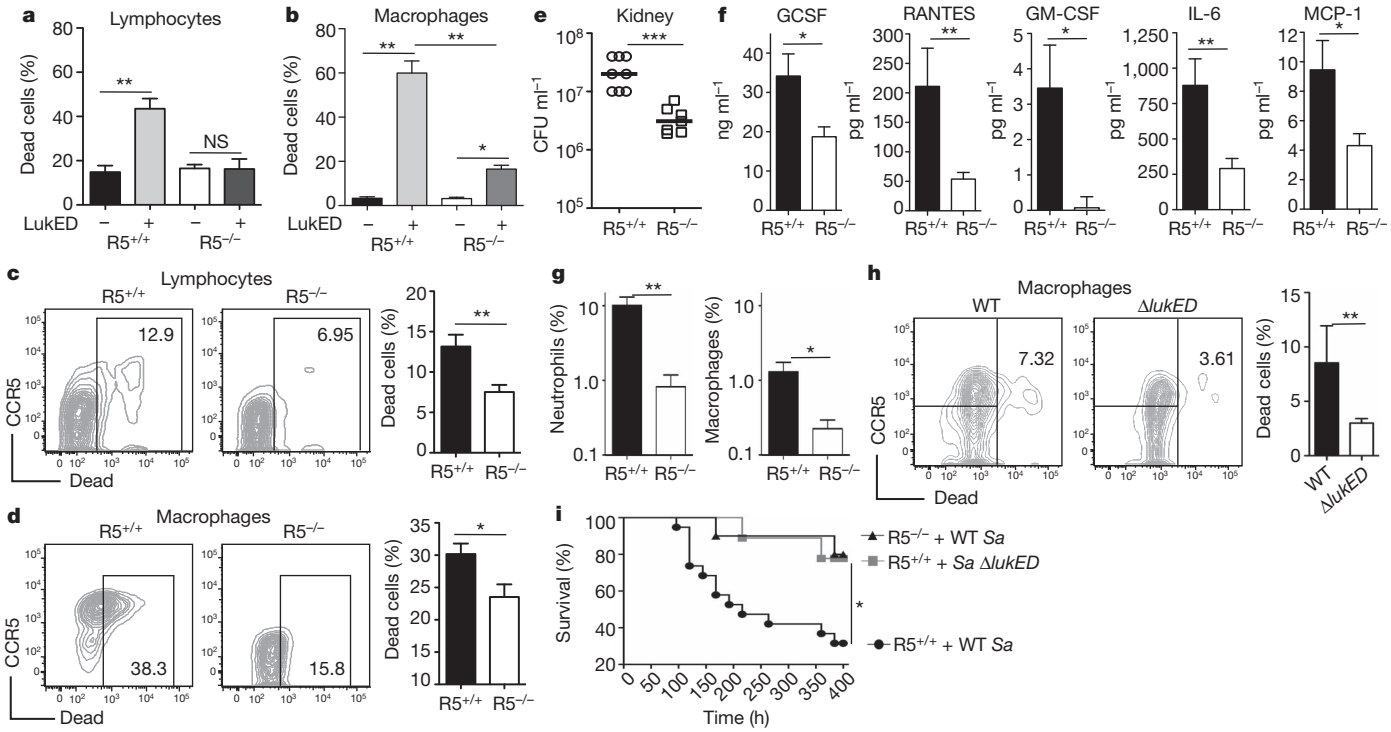


Figure 4 | CCR5$^+$ cell killing is important for *S. aureus* pathogenesis. **a**, **b**, Viability of primary murine peritoneal-elicited immune cells from R5$^{+/+}$ ($n = 3$) or R5$^{-/-}$ ($n = 3$) mice after incubation with LukED (10 μg ml^{-1}). **c**, **d**, *In vivo* viability of recruited immune cells from R5$^{+/+}$ ($n = 10$) or R5$^{-/-}$ ($n = 10$) mice challenged with live *S. aureus* Δ*rot*. **e**, Bacterial colony-forming units (CFU) recovered from the kidneys of R5$^{+/+}$ ($n = 8$) or R5$^{-/-}$ ($n = 9$) mice infected for 96 h with WT *S. aureus*. **f**, Serum cytokine and chemokine amounts from animals in **e**. **g**, Quantification of neutrophils and macrophages recovered from infected kidneys 96 h after infection. **h**, *In vivo* viability of recruited macrophages from R5$^{+/+}$ mice challenged with *S. aureus* WT ($n = 10$) or Δ*lukED* ($n = 10$). **i**, 'Survival' of R5$^{+/+}$ mice infected with WT *S. aureus* ($n = 10$) or a Δ*lukED* mutant ($n = 10$) and R5$^{-/-}$ infected with WT *S. aureus* ($n = 20$). FACS plots show a representative from one of 10 infected animals. *$P < 0.05$; **$P \leq 0.001$; ***$P \leq 0.0001$ by one-way analysis of variance (**a**, **b**), Student's *t*-test (**c**–**h**) and Mantel–Cox test (**i**). Bar graphs, mean ± s.d.

The finding that LukED selectively kills CCR5$^+$ T cells, macrophages and dendritic cells extends the repertoire of immune cells targeted by this leukotoxin and supports a role for these leukocytes in the resolution of *S. aureus* infection. The *lukED* gene is believed to be present in many clinically relevant strains (>70%) including clones responsible for most infections in the USA and Germany, although it is absent in a subset of strains causing hospital-acquired infection (for example EMRSA15/16) in the UK[28–30]. Most isolates lacking *lukED* seem to be of clonal complex 30 (USA200/EMRSA16), which is known to produce low amounts of cytotoxins[31]. Conceivably, the pathogenesis of these strains is influenced by the weakened immune status of hospitalized patients rather than toxic molecules. In contrast, we predict that virulent clinical strains producing large amounts of LukED (for example, clonal complex 8)[7] use the toxin to eliminate antigen-presenting cells as well as *S. aureus*-specific CCR5$^+$ Th1/Th17 cells, which are induced by the bacterium[32] and are protective against infection[33,34]. In support of this hypothesis, we demonstrate that LukED kills CCR5$^+$ cells *in vivo* during systemic infection and that mice lacking CCR5 are protected from the mortality associated with acute *S. aureus* disease. Current systemic murine infection models are insufficient to evaluate reliably CCR5^{hi} T-cell susceptibility to LukED (data not shown). However, our *in vitro* data and *in vivo* studies with CCR5$^+$ macrophages strongly support the notion that subsets of CCR5^{hi} T cells are also targeted *in vivo*.

Interestingly, LukED-mediated toxicity towards neutrophils and monocytes is not blocked by maraviroc (data not shown), suggesting LukED targets these cells through alternative and non-redundant mechanisms. This point also implies a role for CCR5$^+$ myeloid cells and T cells in resolving acute infection, one that extends beyond the initial control of infection imparted by neutrophils. The finding that LukED toxicity towards CCR5$^+$ cells is potently neutralized by a clinically approved CCR5 antagonist (maraviroc) suggests that these types of drug could provide much-needed therapeutic alternatives in the treatment of *S. aureus* infections.

METHODS SUMMARY

Cell lines and primary human cells were maintained in RPMI plus 10% fetal bovine serum with penicillin and streptomycin, unless supplemented as otherwise indicated, and were incubated with LukE, LukD or LukED as previously described[7]. All blood samples were obtained from anonymous healthy donors as buffy coats (New York Blood Center). The New York Blood Center obtained written informed consent from all participants. Animal experiments were performed in accordance with the Institutional Animal Care and Use Committee at New York University School of Medicine. CCR5-overexpressing cell lines and *CCR5* shRNA knockdowns were generated by lentiviral-based tranduction as previously described[35]. Isolation of human peripheral blood mononuclear cells (PBMC) and their sorted subsets was performed as previously described[19].

Full Methods and any associated references are available in the online version of the paper.

Received 28 March; accepted 26 October 2012.
Published online 12 December 2012.

1. Foster, T. J. Immune evasion by staphylococci. *Nature Rev. Microbiol.* **3,** 948–958 (2005).
2. Bischofberger, M., Iacovache, I. & Gisou van der Goot, F. Pathogenic pore-forming proteins: function and host response. *Cell Host Microbe* **12,** 266–275 (2012).
3. Menestrina, G. *et al.* Ion channels and bacterial infection: the case of β-barrel pore-forming protein toxins of *Staphylococcus aureus*. *FEBS Lett.* **552,** 54–60 (2003).
4. Dumont, A. L. *et al.* Characterization of a new cytotoxin that contributes to *Staphylococcus aureus* pathogenesis. *Mol. Microbiol.* **79,** 814–825 (2011).

5. Van de Velde, H. Etude sur le mécanisme de la virulence du staphylocoque pyogène. *Cellule* **10,** 403–460 (1894).
6. Panton, P. N. & Valentine, F. C. O. Staphylococcal toxin. *Lancet* **i,** 506–508 (1932).
7. Alonzo, F. III *et al. Staphylococcus aureus* leucocidin ED contributes to systemic infection by targeting neutrophils and promoting bacterial growth *in vivo*. *Mol. Microbiol.* **83,** 423–435 (2012).
8. Morner, A. *et al.* Primary human immunodeficiency virus type 2 (HIV-2) isolates, like HIV-1 isolates, frequently use CCR5 but show promiscuity in coreceptor usage. *J. Virol.* **73,** 2343–2349 (1999).
9. Deng, H. *et al.* Identification of a major co-receptor for primary isolates of HIV-1. *Nature* **381,** 661–666 (1996).
10. Doranz, B. J. *et al.* A dual-tropic primary HIV-1 isolate that uses fusin and the beta-chemokine receptors CKR-5, CKR-3, and CKR-2b as fusion cofactors. *Cell* **85,** 1149–1158 (1996).
11. Didigu, C. A. & Doms, R. W. Novel approaches to inhibit HIV entry. *Viruses* **4,** 309–324 (2012).
12. Strizki, J. M. *et al.* Discovery and characterization of vicriviroc (SCH 417690), a CCR5 antagonist with potent activity against human immunodeficiency virus type 1. *Antimicrob. Agents Chemother.* **49,** 4911–4919 (2005).
13. Baba, M. *et al.* A small-molecule, nonpeptide CCR5 antagonist with highly potent and selective anti-HIV-1 activity. *Proc. Natl Acad. Sci. USA* **96,** 5698–5703 (1999).
14. Lee, B. *et al.* Epitope mapping of CCR5 reveals multiple conformational states and distinct but overlapping structures involved in chemokine and coreceptor function. *J. Biol. Chem.* **274,** 9617–9626 (1999).
15. Rich, R. L., Miles, A. R., Gale, B. K. & Myszka, D. G. Detergent screening of a G-protein-coupled receptor using serial and array biosensor technologies. *Anal. Biochem.* **386,** 98–104 (2009).
16. Liu, R. *et al.* Homozygous defect in HIV-1 coreceptor accounts for resistance of some multiply-exposed individuals to HIV-1 infection. *Cell* **86,** 367–377 (1996).
17. Samson, M. *et al.* Resistance to HIV-1 infection in Caucasian individuals bearing mutant alleles of the CCR-5 chemokine receptor gene. *Nature* **382,** 722–725 (1996).
18. El Hed, A. *et al.* Susceptibility of human Th17 cells to human immunodeficiency virus and their perturbation during infection. *J. Infect. Dis.* **201,** 843–854 (2010).
19. Wan, Q. *et al.* Cytokine signals through PI-3 kinase pathway modulate Th17 cytokine production by CCR6+ human memory T cells. *J. Exp. Med.* **208,** 1875–1887 (2011).
20. Saita, Y., Kondo, M. & Shimizu, Y. Species selectivity of small-molecular antagonists for the CCR5 chemokine receptor. *Int. Immunopharmacol.* **7,** 1528–1534 (2007).
21. Golding, H. *et al.* Inhibition of HIV-1 infection by a CCR5-binding cyclophilin from *Toxoplasma gondii*. *Blood* **102,** 3280–3286 (2003).
22. Aliberti, J. *et al.* Molecular mimicry of a CCR5 binding-domain in the microbial activation of dendritic cells. *Nat. Immunol.* **4,** 485–490 (2003).
23. Rahbar, R., Murooka, T. T. & Fish, E. N. Role for CCR5 in dissemination of vaccinia virus *in vivo*. *J. Virol.* **83,** 2226–2236 (2009).
24. Lalani, A. S. *et al.* Use of chemokine receptors by poxviruses. *Science* **286,** 1968–1971 (1999).
25. Hummel, S., Schmidt, D., Kremeyer, B., Herrmann, B. & Oppermann, M. Detection of the *CCR5-Δ32* HIV resistance gene in Bronze Age skeletons. *Genes Immun.* **6,** 371–374 (2005).
26. Lucotte, G. Frequencies of 32 base pair deletion of the (*Δ32*) allele of the CCR5 HIV-1 co-receptor gene in Caucasians: a comparative analysis. *Infect. Genet. Evol.* **1,** 201–205 (2002).
27. Hedrick, P. W. & Verrelli, B. C. 'Ground truth' for selection on *CCR5-Δ32*. *Trends Genet.* **22,** 293–296 (2006).
28. Moore, P. C. & Lindsay, J. A. Molecular characterisation of the dominant UK methicillin-resistant *Staphylococcus aureus* strains, EMRSA-15 and EMRSA-16. *J. Med. Microbiol.* **51,** 516–521 (2002).
29. Vandenesch, F. *et al.* Community-acquired methicillin-resistant *Staphylococcus aureus* carrying Panton-Valentine leukocidin genes: worldwide emergence. *Emerg. Infect. Dis.* **9,** 978–984 (2003).
30. von Eiff, C., Friedrich, A. W., Peters, G. & Becker, K. Prevalence of genes encoding for members of the staphylococcal leukotoxin family among clinical isolates of *Staphylococcus aureus*. *Diagn. Microbiol. Infect. Dis.* **49,** 157–162 (2004).
31. DeLeo, F. R. *et al.* Molecular differentiation of historic phage-type 80/81 and contemporary epidemic *Staphylococcus aureus*. *Proc Natl Acad Sci USA* **108,** 18091–18096 (2011).
32. Zielinski, C. E. *et al.* Pathogen-induced human T_H17 cells produce IFN-γ or IL-10 and are regulated by IL-1β. *Nature* **484,** 514–518 (2012).
33. Cho, J. S. *et al.* IL-17 is essential for host defense against cutaneous *Staphylococcus aureus* infection in mice. *J. Clin. Invest.* **120,** 1762–1773 (2010).
34. Lin, L. *et al.* Th1-Th17 cells mediate protective adaptive immunity against *Staphylococcus aureus* and *Candida albicans* infection in mice. *PLoS Pathog.* **5,** e1000703 (2009).
35. Oswald-Richter, K. *et al.* Identification of a CCR5-expressing T cell subset that is resistant to R5-tropic HIV infection. *PLoS Pathog.* **3,** e58 (2007).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank members of the Torres laboratory, D. R. Littman, M. Lu, and A. Darwin for reading this manuscript. We also thank V. KewalRamani for providing reagents, and S. Polsky for assistance with purification of PBMCs. This research was supported by New York University School of Medicine Development Funds, an American Heart Association Scientist Development Grant (09SDG2060036) to V.J.T. and National Institutes of Health (NIH) grants R56-AI091856-01A1 to V.J.T., NIH training grant T32-AI007180 to F.A., A.L.D. and S.A.R., NIH R42-MH084372-02A1 to D.G.M., and NIH R21-AI087973 and R01-AI065303 grants to D.U.

Author Contributions F.A. and V.J.T. identified CCR5 as the LukED receptor. F.A., A.L.D. and T.R.-R. purified the toxins. S.A.R. generated the *CCR5* shRNA knockdown and CCR5 over-expressing cells. F.A., S.A.R. and A.L.D. performed the cytotoxicity assays of cell lines. L.K. purified and sorted primary cells. D.U. designed the experiments for the effect of LukED on human cells. L.K. performed the experiments with primary human cells and S.A.R. performed the HIV infection experiments. F.A. and T.R.-R. conducted the biochemical and cell binding studies with LukED and GFP fusion proteins. F.A. and T.R.-R. conducted the animal studies. D.M. performed the surface plasmon resonance experiments. N.R.L. provided cDNA plasmids and the Δ32 CCR5 primary cells. V.J.T. and D.U. coordinated and directed the project. All authors discussed the data and commented on the manuscript. F.A., D.U. and V.J.T. interpreted the data and wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to V.J.T. (victor.torres@nyumc.org) or D.U. (derya.unutmaz@nyumc.org).

METHODS

Cell culture conditions and viruses. Mammalian cells were maintained at 37 °C with 5% CO_2 in RPMI supplemented with 10% fetal bovine serum (FBS; Atlanta Biologicals) and penicillin (100 U ml^{-1}) and streptomycin (0.1 mg ml^{-1}) (Mediatech) unless stated otherwise. Lentivirus-based overexpression and knockdown of human CCR5 were conducted according to previously described transduction methods[19]. Virus stocks were produced by DNA transfection mediated by calcium phosphate as described[35]. CCR5 overexpressing and shRNA-encoding viruses, including non-coding shRNA or HSA (mCD24)-overexpressing controls, were used at a multiplicity of infection of 1–3. HIV-R5 virus used for infection of primary T cells was used at a multiplicity of infection of 0.3.

Isolation of human PBMC, T-cell purification and activation. Blood was obtained from de-identified, consenting healthy adult donors as Buffy coats (New York Blood Center) and from Δ32/Δ32 CCR5 donors. Human peripheral blood mononuclear cells (PBMCs) were isolated from blood using a Ficoll-Paque PLUS (GE Amersham) gradient. Resting CD4$^+$ and CD8$^+$ human T cells were purified as previously described[19]. Briefly, CD4$^+$ and CD8$^+$ T cells were isolated from purified PBMCs using Dynal CD4$^+$ or CD8$^+$ Isolation Kits (Life Technologies) and were more than 99% pure. To purify naive, central memory and effector memory subsets, isolated CD4$^+$ and CD8$^+$ cells were stained with CCR7 and CD45RO antibodies, and CD45RO$^-$CCR7$^+$ (TN), CD45RO$^+$CCR7$^+$ (TCM), CCR7$^-$ (effector memory T lymphocyte) subsets were sorted using a flow cytometer (FACSAria; BD Biosciences). In some experiments, total CD45RO$^+$ (T_M) cells were sorted into CCR5$^+$ and CCR5$^-$ subsets. Sorted subsets were more than 98% pure. Primary human CD4$^+$ T cells for HIV-R5 infections were activated using anti-CD3/CD28 coated beads (Dynabeads, Invitrogen) and maintained in RPMI + penicillin and streptomycin + 10% FBS supplemented with 200 U ml^{-1} IL-2 and 2 mM L-glutamine (Mediatech). In some experiments, CD4$^+$ T cells were cultured in 20 ng ml^{-1} IL-7 plus IL-15 (R&D Systems) for 7 days. All experiments with primary PBMCs from WT CCR5 donors were performed with cells from at least three independent donors. Experiments using Δ32 CCR5 PBMCs were performed with cells from two donors.

Generation of primary human monocyte-derived macrophages, and dendritic cells. Monocyte-derived macrophages and dendritic cells from healthy donors were generated from CD14$^+$ cells as previously described[35]. Monocyte (CD14$^+$) cells were isolated from PBMCs using anti-CD14 antibody-coated bead-based sorting using AutoMACS (Miltenyi Biotec) and were typically more than 99% pure. Monocyte-derived macrophages were generated from CD14$^+$ cells by supplementing the culture medium with human granulocyte macrophage colony-stimulating factor (50 ng ml^{-1})[36]. Monocyte-derived dendritic cells were generated from CD14$^+$ cells by supplementing the culture medium with human granulocyte macrophage colony-stimulating factor (50 ng ml^{-1}) + IL-4 (40 ng ml^{-1})[37]. Cells were cultured for 5 days in the differentiation condition, followed by addition of LukED as already described.

CCR5 ligands and inhibitors. Maraviroc and TAK-779 were obtained through the AIDS Research and Reference Reagent Program, Division of AIDS, NIAID, NIH. Vicriviroc was purchased from Selleck Chemicals. Recombinant human Rantes (CCL-5) and macrophage inflammatory protein-1β (MIP-1β, CCL-4) were obtained from R&D Systems. Macrophage inflammatory protein 1α (MIP-1α, CCL-3) was obtained from Biolegend. Maraviroc was used at 100 ng ml^{-1} unless otherwise indicated.

FACS analysis. Cells were stained as previously described[35]. For intracellular staining, CD4$^+$ T-cell cultures were stimulated for 5 h at 37 °C with PMA, ionomycin and Golgistop (BD Biosciences). Stimulated cells were washed with PBS and stained with Fixable Viability Dye to gate on live cells. Cells were then fixed and permeabilized by a commercially available intracellular staining kit (eBioscience) according to the manufacturer's protocol. All FACS data were acquired on an LSRII flow cytometer (BD Biosciences) using FACSDiva software. Data were analysed using Flowjo software (Treestar).

Antibodies and dyes. Antibodies used for surface and intracellular staining of primary human cells included the following: CD3-Percp Cy5.5 (clone UCHT1), CD4-Alexa700 (clone OKT4), CD8-Pacific Blue (clone RPA-T8), CXCR3-Percp Cy5.5 (clone G025H7), IL-17-Alexa488 (clone BL168), IFN-γ-Alexa700 (clone 4S.B3) (Biolegend), CD45RO-PeCy7 (clone UCHL1), CCR6-biotin (clone 11A9), CCR4-PE (clone 1G1), CCR5-PE (clone 2D7) or CCR5-APC-Cy7 (clone 2D7), streptavidin-APC, HSA-PE (clone M1/69) (BD Biosciences), and CCR7-FITC (clone 150503) (R&D systems), IL-22- PerCP-eFluor710 (clone 22URTI) (eBioscience).

Antibodies used for surface staining of primary murine cells included the following: CD3ε-APC (clone 145-2C11), CD11b-PeCy7 (clone M1/70), CD11b-FITC (clone M1/70) Ly6G-FITC (clone 1A8), Ly6G-PE (clone 1A8), CCR5-biotin (C34-3448), CD16/CD32 Fc Block (clone 2.4G2) (BD Biosciences), F4/80-APC (clone BM8), F4/80-PeCy7 (clone BM8) streptavidin-PerCP.Cy5.5, and B220-A700 (clone RA3-6B2) (Biolegend). Fixable viability dyes eFluor-450 and eFluor-780 were obtained from eBioscience.

Antibodies used for LukE-CCR5 interaction mapping included the following: CCR5 clones 45533, 45529, 45531, 45517, 45523, 45549, 3A9, 45502 and 45519 (ref. 14) (R&D systems). The control CXCR4 antibody used in these studies was clone 44716 (R&D systems).

Leukotoxin treatments. Jurkat, H9, Hut-R5 and GHOST cell lines, primary human PBMCs and their sorted subsets, as well as primary murine peritoneal-elicited cells, were incubated with LukE, LukD or LukED as previously described[7]. In all experiments cells were seeded into a 96-well plate (1×10^5 to 2×10^5 cells per well), treated for 1 h at 37 °C and evaluated for morphological changes and ethidium bromide (EtBr) uptake by microscopy, or viability using CellTiter (Promega), CytotoxOne (Promega), cell scatter by FACS and staining with commercial viability dyes (eBioscience). CellTiter, CytotoxOne and EtBr measurements were made using an EnVision 2103 Plate Reader (Perkin-Elmer). Intoxications were done in the presence of specific inhibitors (maraviroc, TAK-779 and vicriviroc), chemokines (CCL3, CCL4, CCL5) or monoclonal CCR5 and CXCR4 antibodies where indicated in the text.

***S. aureus in vitro* infection experiments.** *S. aureus* (Newman) Δ*rot*, and Δ*rot* Δ*lukED*[7], were subcultured for 5 h in tryptic soy broth followed by washing in RPMI plus 10% FBS and normalization to 1×10^9 CFU per millilitre in this same media. Normalized bacteria were then added to 2×10^5 Jurkat and Jurkat-R5 cells (multiplicity of infection 10:1) that had been pre-stained with α-CCR5-PE antibody (clone 2D7) and mixed at a ratio of 50:50. Staining of CCR5 with α-CCR5-PE antibody (clone 2D7) was previously determined to be stable for longer than 6 h on the surface of Jurkat-R5 cells yet did not influence the killing of these cells by LukED (data not shown and Supplementary Fig. 13). Infected cells were incubated at 37 °C + 5% CO_2 for 4 h followed by the addition of lysostaphin to kill all bacteria. Samples were then analysed on a BD LSRII flow cytometer. Depletion of CCR5$^+$ compared with CCR5$^-$ cells was evaluated and shown graphically as the percentage of dead cells relative to controls with no toxin. For studies with maraviroc, the inhibitor was added to cells 30 min before the addition of bacteria as described above. Experiments were conducted three times in triplicate.

Generation of GFP fusion proteins. To generate recombinant N-terminal His_6–GFP-tagged LukE and LukD, the mature protein coding sequences of LukE and LukD from *S. aureus* Newman genomic DNA were PCR-amplified using the following primers: *lukE*-F-SalI (5′-CCCC-GTCGAC-AATACTAA TATTGAAAAT-3′), *lukD*-F-SalI (5′-CCCC-GTCGAC-GCTCAACATATCA CA-3′), *lukE*-R-NotI (5′-CCCC-GCGGCCGC-tta-ATTATGTCCTTTCACTT TAATTTCGTG-3′) and *lukD*-R-NotI (5′-CCCC-GCGGCCGC-tta-TACTCC AGGATTAGTTTCTTTAGAATC-3′). Amplified sequences were subcloned into pET-41b (Novagen), resulting in a fusion of His_6–GFP with the N terminus of mature LukE or LukD. Recombinant plasmids were transformed into *Escherichia coli* DH5α and transformants selected by kanamycin resistance. Positive clones were transformed into *E. coli* LysY/LacQ (New England BioLabs) for protein expression and purification.

Leukotoxin purification. LukE, LukD, GFP–LukE, GFP–LukD, LukS, LukF, LukA and LukB were purified from *E. coli* LysY/LaqQ as previously described[4,7] followed by endotoxin removal with Detoxi-Gel Endotoxin Removal Gel (Thermo Scientific). The following alterations were made for purification of recombinant GFP–LukE and GFP–LukD. Upon sonication of bacterial cell pellets, lysates were incubated with 1% Triton X-100 for 1 h at room temperature. After incubation, lysates were centrifuged for 60 min at 12,350*g* and passed through a 0.22 μm filter before completing the purification protocol as described[7].

LukED membrane association studies. Association of LukED with the surface of CCR5$^+$ cells was measured as follows. A toxic dose of purified recombinant GFP–LukE or GFP–LukD with LukD or LukE, respectively, (final concentration 10 μg ml^{-1}) was incubated for 30 min on ice with sorted CD4$^+$CCR5$^+$ or CD4$^+$CCR5$^-$ T cells (5×10^4 cells per well) from three independent donors. Cells were gated as GFP positive compared with baseline fluorescence of untreated cells. A total of 50,000 events were collected in all conditions tested. Owing to the high amount of background fluorescence of GFP toxins with the membranes of transduced cell lines, we were unable to use these cells for membrane association assays (data not shown). As an alternative, we used primary CD4$^+$ T cells for membrane association studies. To increase the abundance of CCR5 on these cells and foster reproducible measures of membrane association, CD4$^+$CCR5$^+$ cells were generated from CD4$^+$ cells infected with a lentivirus encoding CCR5 and sorted by FACS as CCR5$^+$ from the resulting CD4$^+$ population after surface staining for CCR5 using 2D7 clone (PE). CD4$^+$CCR5$^-$ cells were sorted from the same population as those cells with undetectable CCR5 surface expression. CCR5 surface staining with 2D7 antibody does not influence toxin killing kinetics and therefore is unlikely to adversely influence membrane association, as the latter is required for the former (Supplementary Fig. 13).

Paradoxically, clone 2D7 also binds to ECL-2 of CCR5 similar to that of clone 45531, which blocks toxin activity. However, 2D7 and 45531 do bind to distinct portions of ECL-2 (the N-terminal portion and carboxy (C)-terminal portion, respectively) perhaps explaining this phenomenon[38]. Alternatively, our staining protocols may not have sufficiently saturated all receptor sites, thereby allowing functional characterization of toxin in the presence of 2D7.

Experiments assessing maraviroc, natural ligand or antibody inhibition of LukED membrane association were conducted in a similar fashion. However, in these instances cells were first pre-incubated for 30 min with maraviroc (100 ng ml^{-1}), CCL5 (5 μg ml^{-1}), 3A9, 45531 or CXCR4 monoclonal antibodies (25 μg ml^{-1}) or buffer before addition of a lethal concentration of LukE–GFP + LukD to the cells (5–10 μg ml^{-1}). After treatment, cells were washed, re-suspended in fixation buffer (FACS buffer + 2% paraformaldehyde) for 15 min at room temperature, washed again, re-suspended in FACS buffer, and the fluorescence of bound toxin was monitored by flow cytometry. Cells are shown as the percentage that were GFP positive.

Surface plasmon resonance analysis of LukE and LukD binding to solubilized CCR5 and CXCR4. Binding kinetics of LukE and LukD to CCR5 and CXCR4 by surface plasmon resonance were measured as previously described[15,39–42]. This approach has also been used to detect ligand interactions with CXCR1 and CXCR2 (refs 43, 44). A C9-tagged CCR5 was solubilized using 50 mM HEPES, pH 7.0, 150 mM NaCl, 0.1% DDM, 0.1% CHAPS, 0.02% CHS[15]. This solubilization scheme is known to retain conformationally specific antibody binding to both CCR5 and CXCR4 (ref. 15). Approximately 700 relative units (RU) of the CCR5 receptor was captured onto a 1D4 antibody-bound CM5 chip[15,40,41]. Cells expressing a C9-tagged CXCR4 receptor were also solubilized as a control surface in the same buffer[41]. C9-CXCR4 was captured to approximately 1,200 RU. LukE or LukD was diluted to 1.7 μM in running buffer containing 50 mM HEPES, pH 7.0, 150 mM NaCl, 0.02% CHS, 0.1% DDM and 0.1% Chaps and tested for binding in a threefold dilution series at a flow rate of 50 μl min^{-1}. Each concentration series was replicated twice as shown by the overlaid sensorgrams. All data were collected at 25 °C and conducted at least twice in duplicate.

Biochemical studies to detect interactions between LukED and CCR5. 293T cells were transfected with a vector containing HA-tagged CCR5 (Missouri S&T cDNA Resource Center; www.cdna.org), followed by solubilization (approximately 2.0×10^7 cells per condition) in PBS + 1% Brij010 + Complete EDTA-free protease inhibitor cocktail (Roche). Solubilized CCR5 was then added to 25 μl of nickel resin containing no toxin or bound LukE, LukD or LukED. For the maraviroc, natural ligand and antibody inhibition experiments, the solubilized CCR5 was pre-incubated for 30 min at room temperature with 5 μg ml^{-1} of maraviroc, 10 μg ml^{-1} of each chemokine or 35 μg ml^{-1} of each antibody followed by incubation with nickel resin containing LukE. After incubation with cell lysates, the resin/protein complexes were fixed with 2 mM DTSSP (Pierce) for 30 min, quenched with 20 mM Tris pH 8.0 for 15 min, washed four times in PBS + 1% Brij010 and boiled in 4× SDS boiling buffer. All samples were run on a 10% SDS–PAGE gel at 80 V, followed by transfer to nitrocellulose at 1 A for 1 h. Membranes were blocked in PBS + 0.01% Tween + 5% milk for 1 h and incubated overnight with either α-HA antibody for CCR5 (Covance) or α-His antibody (Cell Sciences) for LukE and LukD. The following day, secondary goat α-mouse-HRP antibody (Bio-Rad) was added to the membranes for 1 h followed by the addition of SuperSignal West Femto Maximum Sensitivity Substrate (Thermo Fisher Scientific) for detection.

Measurement of CCR5 activation by calcium mobilization. CCR5 activation by calcium mobilization in cell lines and primary cells was assessed using the commercial dye Fluo4-AM (Invitrogen). Cells were labelled for 30 min at room temperature with 3 μM Fluo4 in Hanks' balanced salt solution, followed by three washes in Hanks' balanced salt solution and incubation at 37 °C for 30 min. Cells were analysed on a flow cytometer over time and, at 100 s, ligand (CCL3, CCL4, CCL5, 10 ng ml^{-1}) or LukE (10–20 μg ml^{-1}) was added to the cells. Fluorescence was monitored thereafter by flow cytometry (500 events were collected per second) until the indicated completion of each experiment. For conditions in which inhibition of receptor activation was monitored, cells were pre-incubated with either maraviroc (1 μg ml^{-1}) or LukE (10–20 μg ml^{-1}) during the 30 min incubation at 37 °C described above. Graphs show the mean fluorescence of all events collected in 5 s intervals.

Murine *in vitro* and *in vivo* experiments. *In vitro* assessment of peritoneal-elicited immune cell killing by LukED was conducted as follows. Female age-matched (4–6 weeks) C57BL/6 WT or CCR5$^{-/-}$ mice (Taconic) were injected with 1×10^7 CFU of heat-killed *S. aureus* Newman Δ*lukED* intraperitoneally. Twenty-four hours later, mice were injected with an additional 1×10^7 CFU of the same strain. After another 24 h, mice were killed and peritoneal-elicited immune cells were lavaged with 7 ml of PBS followed by lysis of red blood cells in ACK lysing buffer and re-suspension in RPMI + 10% FBS. LukED was then added to cells as described above and incubated for 1 h at 37 °C with 5% CO_2. After incubation, cells were washed in PBS and stained with viability dye followed by surface staining for B220, CD11b, F480, Ly6G, CD3 and CCR5. The percentages of dead cells shown are an average of cells isolated and intoxicated from three independent mice. Means and standard deviations are shown.

For experiments designed to measure *S. aureus* killing of CCR5$^+$ cells *in vivo*, female age-matched (4–6 weeks) C57BL/6 WT or CCR5$^{-/-}$ mice (Taconic) were injected on day 1 with 1×10^7 heat-killed *S. aureus* to promote the recruitment of CCR5$^+$ macrophages and lymphocytes to the peritoneum. On day 2, mice were challenged with live *S. aureus* Δ*rot* followed by the isolation of peritoneal immune cells 16–20 h later. Isolated cells were processed for FACS as described above and the viability of lymphocytes and macrophages was evaluated. The percentages of dead lymphocytes were averaged from 10 WT and 10 CCR5$^{-/-}$ animals; representative FACS plots are shown.

For murine systemic infections, female age-matched (4–6 weeks) C57BL/6 WT or CCR5$^{-/-}$ mice (Taconic) were infected with WT *S. aureus* Newman as previously described[7]. After 96 h, serum was collected and kidneys removed, homogenized, processed for FACS and plated as previously described[7]. All survival curves were conducted as previously described using WT *S. aureus* Newman and an isogenic Δ*lukED* mutant[7]. For flow cytometry of immune cells from WT or Δ*lukED* infected kidneys, organs were removed after 96 h and mechanically homogenized. Immune cells in homogenized tissues were enriched by performing a 40/80 Percoll (GE Healthcare) density gradient centrifugation. Cells were subsequently processed for surface and viability staining thereafter (see above).

36. Gramberg, T., Sunseri, N. & Landau, N. R. Evidence for an activation domain at the amino terminus of simian immunodeficiency virus Vpx. *J. Virol.* **84,** 1387–1396 (2010).
37. Manel, N. *et al.* A cryptic sensor for HIV-1 activates antiviral innate immunity in dendritic cells. *Nature* **467,** 214–217 (2010).
38. Berro, R. *et al.* Multiple CCR5 conformations on the cell surface are used differentially by human immunodeficiency viruses resistant or sensitive to CCR5 inhibitors. *J. Virol.* **85,** 8227–8240 (2011).
39. Stenlund, P., Babcock, G. J., Sodroski, J. & Myszka, D. G. Capture and reconstitution of G protein-coupled receptors on a biosensor surface. *Anal. Biochem.* **316,** 243–250 (2003).
40. Navratilova, I., Sodroski, J. & Myszka, D. G. Solubilization, stabilization, and purification of chemokine receptors using biosensor technology. *Anal. Biochem.* **339,** 271–281 (2005).
41. Navratilova, I., Dioszegi, M. & Myszka, D. G. Analyzing ligand and small molecule binding activity of solubilized GPCRs using biosensor technology. *Anal. Biochem.* **355,** 132–139 (2006).
42. Navratilova, I., Pancera, M., Wyatt, R. T. & Myszka, D. G. A biosensor-based approach toward purification and crystallization of G protein-coupled receptors. *Anal. Biochem.* **353,** 278–283 (2006).
43. Caccuri, F. *et al.* HIV-1 matrix protein p17 promotes angiogenesis via chemokine receptors CXCR1 and CXCR2. *Proc. Natl Acad. Sci. USA* **109,** 14580–14585 (2012).
44. Giagulli, C. *et al.* HIV-1 matrix protein p17 binds to the IL-8 receptor CXCR1 and shows IL-8-like chemokine activity on monocytes through Rho/ROCK activation. *Blood* **119,** 2274–2283 (2012).

ARTICLE

doi:10.1038/nature11801

Structure of a presenilin family intramembrane aspartate protease

Xiaochun Li[1,2]*, Shangyu Dang[1,2]*, Chuangye Yan[1,2], Xinqi Gong[1,2], Jiawei Wang[2,3] & Yigong Shi[1,2]

Presenilin and signal peptide peptidase (SPP) are intramembrane aspartyl proteases that regulate important biological functions in eukaryotes. Mechanistic understanding of presenilin and SPP has been hampered by lack of relevant structural information. Here we report the crystal structure of a presenilin/SPP homologue (PSH) from the archaeon *Methanoculleus marisnigri* JR1. The protease, comprising nine transmembrane segments (TMs), adopts a previously unreported protein fold. The amino-terminal domain, consisting of TM1–6, forms a horseshoe-shaped structure, surrounding TM7–9 of the carboxy-terminal domain. The two catalytic aspartate residues are located on the cytoplasmic side of TM6 and TM7, spatially close to each other and approximately 8 Å into the lipid membrane surface. Water molecules gain constant access to the catalytic aspartates through a large cavity between the amino- and carboxy-terminal domains. Structural analysis reveals insights into the presenilin/SPP family of intramembrane proteases.

Signalling through regulated intramembrane proteolysis (RIP) is ubiquitously conserved from bacteria to humans[1]. The central step of RIP signalling is cleavage of a target protein within the lipid bilayer by a membrane-embedded protease[2,3]. These intramembrane proteases fall into three distinct families: serine protease rhomboid, metalloprotease site-2 protease (S2P), and aspartyl proteases exemplified by presenilin and SPP.

Presenilin is the catalytic component of γ-secretase[4–6], which is responsible for the generation of amyloid-β peptide from amyloid precursor protein (APP)[7]. An altered cleavage pattern of APP, involving aberrant accumulation of a 42-residue amyloid-β peptide (Aβ42) over a 40-residue product (Aβ40), underlies the development of Alzheimer's disease[8,9]. Besides APP, γ-secretase has a number of other important substrates such as Notch[5,6]. Accelerated cleavage of Notch by γ-secretase leads to excessive signalling, contributing to certain cancers[9]. Suppressing the generation of Aβ42 without affecting cleavage of Notch represents an attractive strategy for therapeutic intervention in Alzheimer's disease. SPP is required for degradation of the signal peptide in membrane-targeted proteins. Presenilin and SPP share extensive sequence homology, both with the signature motifs ΦYDΦΦ (in which Φ denotes a hydrophobic residue) on TM6 and ΦGΦGD on TM7 (ref. 10).

In addition to presenilin, the γ-secretase complex contains three other essential components, PEN-2, APH-1 and nicastrin[11]. In this complex, PEN-2 is thought to interact directly with presenilin[12,13], which comprises two auto-proteolysed fragments known as NTF (TM1–6) and CTF (TM7–9)[14]. As the central component, presenilin also associates with APH-1 and nicastrin[14–16]. The central role of presenilin in the γ-secretase complex is manifested by identification of over 150 disease-derived mutations in presenilin[9].

Recent structural investigations of the prokaryotic homologues of rhomboid[17–20] and S2P[21] have facilitated mechanistic understanding of these membrane-embedded proteases. By contrast, detailed structural information on presenilin and SPP has been extremely slow to emerge. The limited structural information includes electron microscopic analysis of the γ-secretase complex[22–24], an NMR analysis of presenilin CTF[25], and a crystal structure of a GXGD peptidase FlaK[26] which shows no sequence homology to presenilin/SPP.

Engineering and crystallization of PSH

We cloned human presenilin 1 (also known as PSEN1, hereafter referred to as PS1) and its homologues in several organisms. Extensive effort aimed at generation of functional recombinant presenilin was only successful at the level of protease activity detection; such recombinant proteins were unfit for biophysical characterization. After years of effort on the eukaryotic presenilin, we sought its archaeal homologues[27]. Of the 13 archaeal presenilin homologues, only two were successfully overexpressed and purified. We focused on PSH (also known as MCMJR1[28]), which gives a higher yield and shares extensive sequence homology with PS1. Unfortunately, the full-length PSH (residues 1–301) exhibits poor behaviour in solution in all non-ionic detergents tested and defied all attempts at crystallization.

We designed a sequence-based protein engineering strategy to improve behaviour of PSH in solution. Sequence alignment between PSH and presenilin homologues from other archaeal species revealed a number of positions that contain highly conserved amino acids in all other archaeal homologues, except PSH. We reasoned that replacement of the non-conserved amino acids in PSH by the conserved ones may improve its behaviour in solution, but is unlikely to affect its folding or protease activity. On the basis of this rationale, we generated 20 PSH variants, each containing one or two point mutations, and individually examined their protease activity using the transmembrane segment of Gurken as the substrate[28] (Supplementary Fig. 1). Only the protease-active PSH variants were selected for evaluation of behaviour in solution and possibly crystallization trials. Single mutations were combined to generate multiple mutations in the same PSH variant (Supplementary Fig. 1). Finally, to increase the likelihood of crystallization, we used V8 protease to remove the flexible surface loops, which resulted in the removal of residues 182–209.

Using this approach, we generated a total of 105 PSH variants for detailed biochemical evaluation. The PSH variant used for final crystallization contains two chains (residues 1–181 and 210–301) and

[1]Ministry of Education Key Laboratory of Protein Science, Center for Structural Biology, School of Life Sciences and School of Medicine, Tsinghua University, Beijing 100084, China. [2]Tsinghua-Peking Joint Center for Life Sciences, Center for Structural Biology, School of Life Sciences and School of Medicine, Tsinghua University, Beijing 100084, China. [3]State Key Laboratory of Bio-membrane and Membrane Biotechnology, Center for Structural Biology, School of Life Sciences and School of Medicine, Tsinghua University, Beijing 100084, China.
*These authors contributed equally to this work.

a total of five missense mutations, D40N, E42S, A147E, V148P and A229V, which mostly affect residues in the surface loops. These sequence modifications have no effect on the protease activity of PSH (Fig. 1a, lanes 1–3). By contrast, mutation of the two catalytic residues Asp 162 and Asp 220 in PSH abolished the protease activity (lanes 4 and 5). Similar to the wild-type PSH, the protease activity of this variant was efficiently inhibited by a known presenilin inhibitor III-31-C (ref. 29; Fig. 1a, lanes 6–14). Intriguingly, the elution volume of the PSH variant corresponds to a molecular mass in excess of 100 kilodaltons (kDa), indicating oligomerization (Supplementary Fig. 2a). After going through about 160,000 conditions, we eventually generated suitable crystals of this PSH variant for X-ray data collection (Supplementary Fig. 2b, c).

Structure of PSH

The PSH variant was crystallized in three space groups. The structure in the *C*222 space group was determined by platinum-based multi-wavelength anomalous dispersion (MAD) and refined at 3.3 Å resolution (Supplementary Tables 1–3 and Supplementary Figs 3–6). The atomic model was confirmed by selenium anomalous peaks (Supplementary Figs 5 and 6). The atomic coordinates of PSH were used to determine the structures in space groups $C222_1$ and *P*2 (Supplementary Table 1). The PSH protein forms a tetrameric complex in the *C*222 crystals (Supplementary Fig. 4). Each asymmetric unit contains one tetrameric PSH complex in $C222_1$ and two complexes in *P*2.

Each PSH molecule comprises nine transmembrane α-helices (Fig. 1b and Supplementary Fig. 7), and is approximately 50 Å in length, 40 Å in width and 40 Å in height. The surface loops connecting TM1–TM2 and TM7–TM8 are disordered in the crystals. The amino-terminal domain (NTD), comprising TM1–6, forms a horseshoe-shaped structure, which partially surrounds TM7–9 of the carboxy-terminal domain (CTD; Fig. 1b). Consequently, TM7 of the CTD is completely surrounded by other TMs. None of the nine TMs is perpendicular to the lipid membrane. TM1–6 are tilted at angles of 15–35 degrees away from the surface normal of lipid membrane, and TM7–9 are tilted to a smaller extent.

Notably, a large hole traverses through the entire protein and is surrounded by TM2, TM3, TM5 and TM7 (Fig. 1b). This hole, formed mainly by hydrophobic residues, is large enough to allow passage of small ions; it is likely plugged up by lipid molecules and/or other interacting protein in cells. The *in vitro* protease activity of PSH is strongly influenced by the choice of phospholipids (data not shown). A cavity from the cytoplasmic side reaches the active site aspartates, allowing unrestricted solvent access. The catalytic residues Asp 162 and Asp 220 in PSH are located at the bottom of the cavity, approximately 8 Å below the lipid membrane surface from the cytoplasmic side (Fig. 1b). Owing to the hole and the cavity, some transmembrane regions of TM2/TM3/TM5/TM6/TM7 are exposed to solvent. This structural finding is consistent with the biochemical observation that a number of membrane-embedded residues in TM6 and TM7 of PS1, including the catalytic residues Asp 257 and Asp 385, are accessible to crosslinking reagents[30,31]. Similar cavities have also been observed in the electron microscopic structure of the γ-secretase complex[22,24].

The topology of PSH is quite different from that of the other two families of intramembrane proteases rhomboid[17–20] and S2P[21] (Fig. 2a, b). Exhaustive search of the protein data bank using DALI[32] revealed

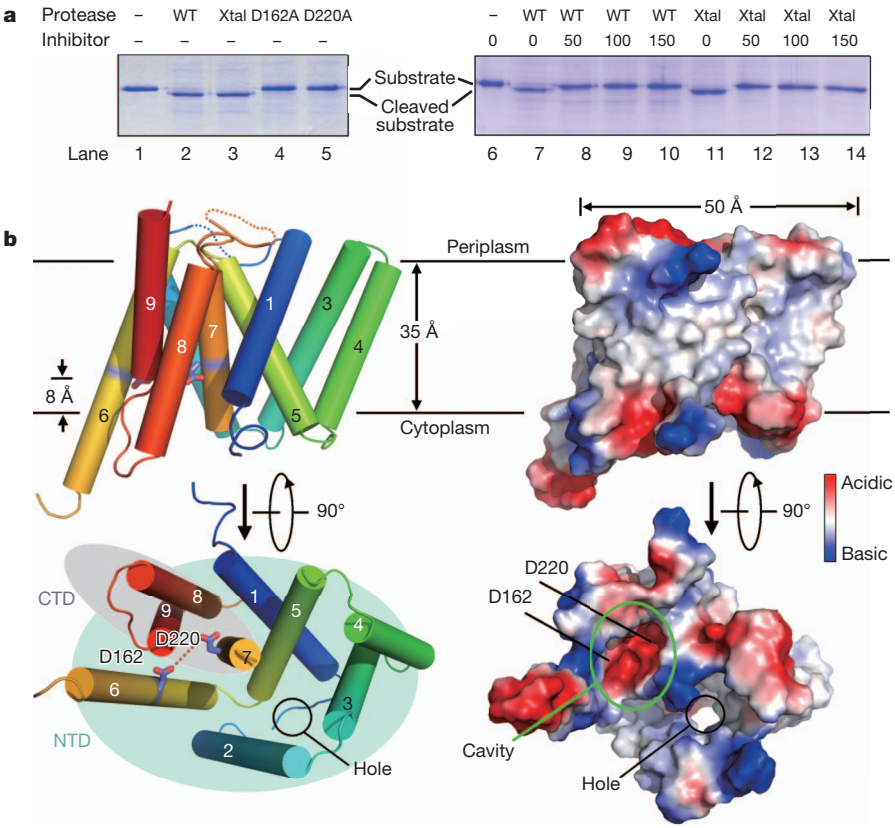


Figure 1 | Overall structure of a presenilin/SPP homologue from the archaeon *Methanoculleus marisnigri* JR1 (PSH). a, The PSH variant used for crystallization (denoted 'Xtal') is catalytically active. As reported[28], the membrane protein Gurken was used as a substrate protein for PSH. The proteolytic activity of the PSH variant (Xtal) is similarly inhibited as the wild-type PSH by III-31-C, an inhibitor known to be specific for aspartyl proteases such as presenilin[29] (right panel). The concentrations of III-31-C shown are in micromolar. **b**, Overall structure of PSH. The nine TMs are divided into the NTD (TM1–6) and CTD (TM7–9). The putative catalytic residues Asp 162 and Asp 220 are located on TM6 and TM7, respectively. The structure is rainbow-coloured, with N terminus in blue and C terminus in red. The top and bottom panels represent two perpendicular views. The structure is shown in cartoon and electrostatic surface representation in the left and right panels, respectively. Figures 1b, 3 and 4a, c were prepared using PyMol[50].

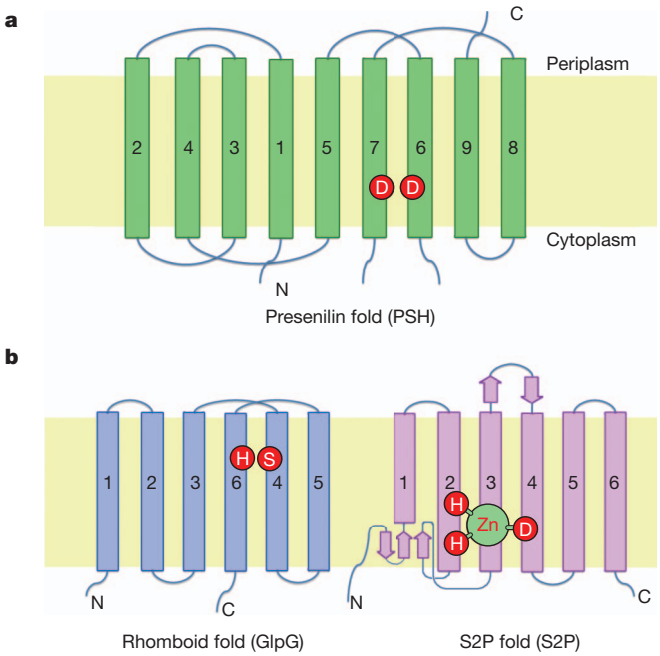


Figure 2 | Topology diagrams of PSH and other families of intramembrane proteases. **a**, Membrane topology diagram of PSH. The two putative catalytic residues, Asp 162 and Asp 220, are indicated. **b**, Membrane topology diagram of the rhomboid protease GlpG from *E. coli* and the metalloprotease S2P from the archaeon *M. jannaschii*. The catalytic dyad residues Ser-His of GlpG and the three zinc-binding residues of S2P are indicated.

no entry that is similar to the structure of PSH over its entire nine TMs. In fact, no entry shows structural similarity to PSH over four TMs. The closest entry (with a similarity *Z*-score of 4.0) is the structure of Sec15 C-terminal domain[33], which only had 61 Cα atoms aligned to PSH with a root-mean-squared deviation (r.m.s.d.) of 3.1 Å. This analysis indicates that PSH may represent a previously unreported protein fold, which we would like to name the presenilin fold (Fig. 2a).

Tetrameric assembly of PSH

The PSH tetramer contains four hydrophobic holes, each traversing through the entire TM region (Supplementary Fig. 8a). Each PSH molecule contacts two neighbouring molecules exclusively through their membrane-spanning regions. Hydrophobic amino acids in TM2 and TM6 of one PSH molecule make van der Waals contacts to residues of the same type in TM3 and TM4 of an adjacent PSH molecule. Notably, four aromatic residues on TM3—Phe 72, Phe 76, Phe 79 and Phe 82—are placed on the same side of the α-helix and interact with four hydrophobic amino acids, Leu 48, Ile 51, Leu 55 and Leu 59, all on TM2 of the neighbouring molecule (Supplementary Fig. 8b). An identical homotetrameric complex of PSH is formed in crystals of the space groups $C222_1$ and $P2$ (Supplementary Fig. 9). As anticipated, the different PSH molecules with the same space group or between different space groups are all nearly identical to one another, with a pair-wise r.m.s.d. of 0.3–0.6 Å over the aligned Cα atoms.

Active site and accessibility

Similar to other aspartyl proteases such as pepsin, PSH also contains two catalytic residues Asp 162 and Asp 220, located in the intramembrane aspartyl protease signature motifs, $VYD_{162}AI$ on TM6 and $MGMGD_{220}$ on TM7 (Fig. 2a ands Supplementary Fig. 7). The reaction mechanism requires placement of the two aspartate side chains to be within hydrogen-bond distance; in pepsin[34], this distance is 3.1 Å. In the structure of PSH, however, Asp 162 and Asp 220 face each other but are separated by a distance of 6.7 Å (Fig. 3a). This observation suggests that substrate binding may trigger a conformational change in PSH that moves the two aspartate residues closer to carry out catalysis. Such an induced-fit mechanism might safeguard PSH, a membrane-embedded protease, against non-specific cleavages. Accordingly, we further speculate that the activated conformation of PSH may be stabilized by the presence of a substrate or inhibitor. Alternatively, although less likely, the extended distance between the two aspartates could be caused by crystal packing or other artefacts of *in vitro* manipulation. Regardless of these scenarios, the observed inactive conformation of PSH is apparently stable under our experimental conditions, and additional conformations of the presenilin-like protease may be captured by future structural investigation.

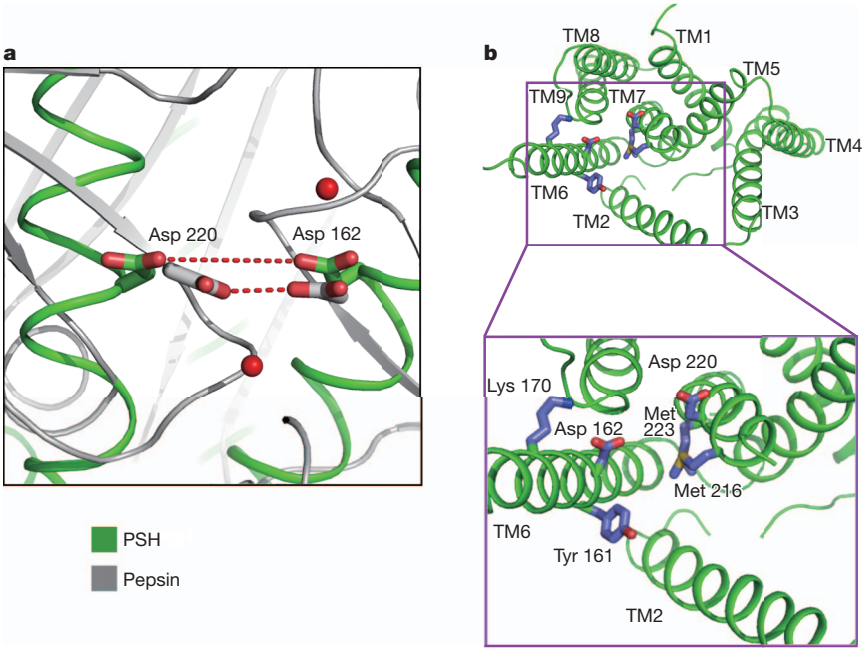


Figure 3 | Conformation of the active site. **a**, A close-up comparison of the active site conformation between PSH and pepsin. The two aspartate residues in PSH and pepsin are separated by distances of 6.7 Å and 3.1 Å, respectively. PSH and pepsin are coloured green and grey, respectively. Two water molecules in pepsin are shown as red spheres. **b**, A close-up view of the active site conformation in PSH. Four residues surrounding Asp 162 and Asp 220 are shown.

Cleavage of the peptide bond requires water. Access of water molecules to the catalytic aspartates in PSH is unrestricted, due to the presence of a large solvent cavity facing the cytoplasm (Fig. 1b). This feature is similar to that of rhomboid, where solvent access to the active site is provided by a cavity that opens to the extracellular side[17–20]. In S2P, however, water access is restricted in the closed conformation and only becomes profuse in the open conformation[21].

Analysis of the TM organization reveals two potential routes for substrate entry, through the open space either between TM6 and TM9 or between TM6 and TM2 (Fig. 3b). The first possibility (between TM6 and TM9) is supported by experimental evidence. For PS1, TM9 is thought to be directly involved in substrate binding[35], and photoaffinity labelling experiments suggest that substrate entry site is between NTF and CTF[36]. Structural consideration also favours this possibility, as Asp 162 and Asp 220 are both unobstructed in the case of substrate entry from between TM6 and TM9 (Supplementary Fig. 10a). By contrast, there is considerable steric hindrance from Tyr 161 on TM6 and Met 216/Met 223 on TM7 if substrate proteins were to gain lateral access to the active site from the space between TM6 and TM2 (Fig. 3b and Supplementary Fig. 10b). This analysis, performed on the inactive conformation of PSH, indicates that substrate proteins are likely to enter the active site through the space between TM6 and TM9. We acknowledge the caveat that such analysis should be ideally carried out for the active protease conformation.

Implication for human presenilin 1

The primary sequences of PSH and PS1 share 19.3% identity and 52.8% similarity (Supplementary Fig. 7). The highly conserved amino acids among PSH, PS1 and their homologues in *Xenopus laevis* and *Drosophila melanogaster* represent 48.2% of the total aligned sequences (Fig. 4a and Supplementary Fig. 7). Importantly, the signature motifs for catalysis are conserved between PSH and PS1. Therefore, the structure of PS1 should be very similar to that of PSH, with observed key structural features in PSH likely to be preserved in PS1. On the basis of sequence alignment, we generated a membrane topology diagram for PS1 (Fig. 4b) and built an energy-minimized atomic

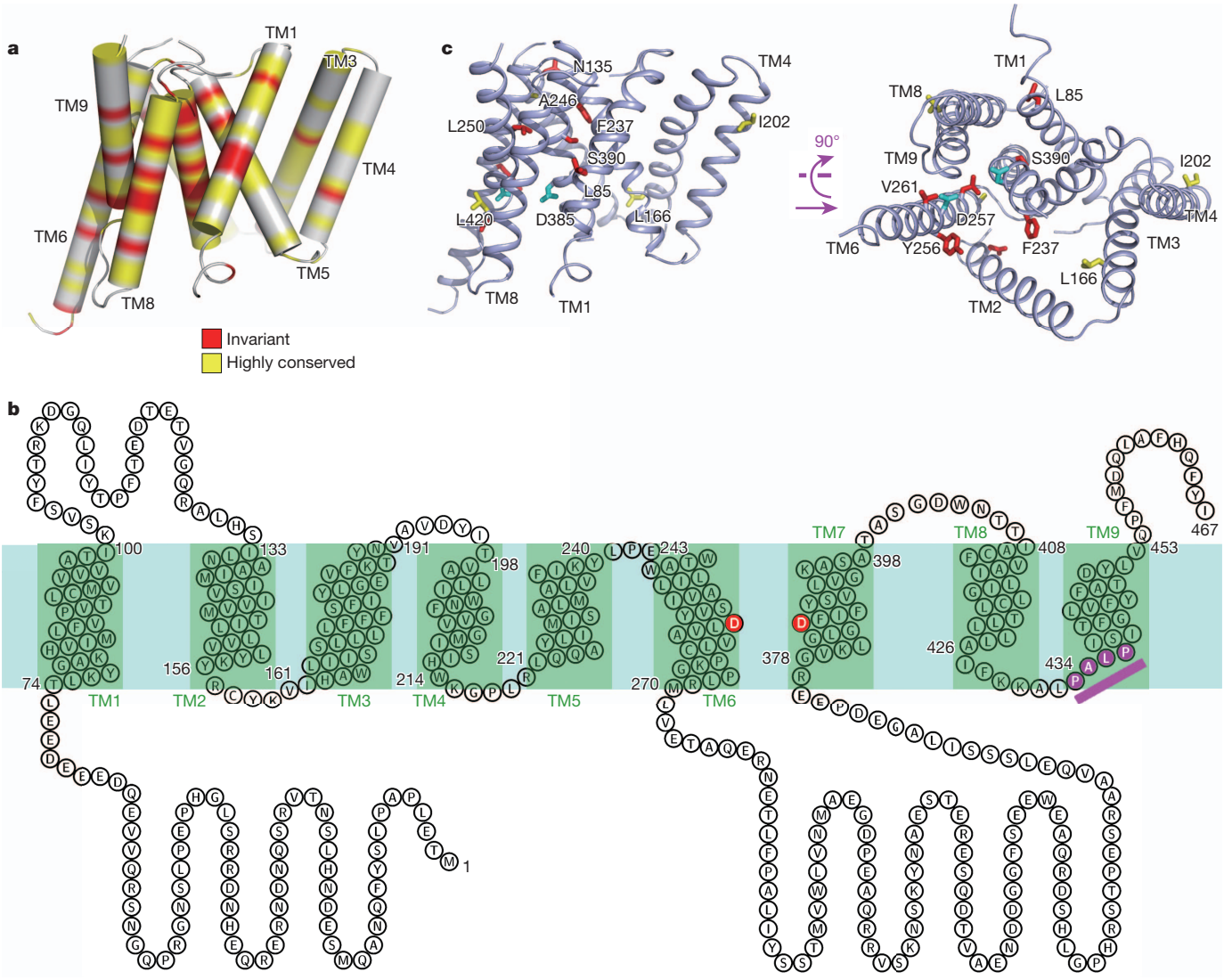


Figure 4 | Homology modelling of human presenilin 1. **a**, Amino acids that constitute the core of the PSH structure are highly conserved among PSH, human PS1, *Xenopus laevis* PS1 and *Drosophila melanogaster* PS1. The invariant and highly conserved amino acids in all four proteins are coloured red and yellow, respectively. **b**, A schematic diagram on the predicted membrane topology of PS1. Demarcations for all nine TMs are labelled. Catalytic aspartates are highlighted in red. Substrate recognition sequence PALP is shown in magenta. **c**, A structure-based model of the transmembrane core of human PS1. The atomic model of human PS1 transmembrane core was generated as described in Methods. Two perpendicular views are shown. Among the dozens of amino acids in PS1 that were targeted for mutations in neurodegenerative diseases, 12 representative residues, which are either invariant (red) or highly conserved (yellow), are shown here.

model for PS1 (Fig. 4c). Comparison of our PS1 model with the NMR structure of PS1 CTF[25] reveals a shared overall topology but considerable differences in TM7 and TM9 (Supplementary Fig. 11). Notably, the catalytic residue Asp 385 in the NMR structure, but not in our PS1 model, points to the opposite direction of the other catalytic Asp 257 on TM6.

Greater than 50% of the disease-derived mutations in PS1 target highly conserved amino acids among PSH and PS1 homologues (Supplementary Fig. 7). We selected 12 representative mutations, which affect 8 invariant and 4 conserved amino acids between PSH and PS1, and mapped the affected residues onto the PS1 model (Fig. 4c). Next, we generated the corresponding mutations in PSH and examined the protease activities of these variants (Supplementary Fig. 12). These mutations are classified into three categories. The first category, including two mutations L108F and F142I (Supplementary Fig. 12, lanes 10–13), has little effect on protease activity. The second category decreases, but does not abolish, the protease activity of PSH. It includes eight mutations in PSH: L17P, N46D, I77P, V151E, L155S, G219A, S225I and V261R (Supplementary Fig. 12, lanes 4–9, 14–20 and 25–30). The third category, including Y161S and V166F, completely abrogated the protease activity of PSH (Supplementary Fig. 12, lanes 21–24).

The effect of the majority of these mutations can be rationalized on the basis of PSH structure. For example, the first category mutations L108F and F142I, affecting residues in TM4 and TM5 of PSH, are far away from the active site and thus are unlikely to be involved in substrate binding or catalysis. The third category mutations Y161S and V166F affect two amino acids in the immediate vicinity of the catalytic residue Asp 162, probably causing pronounced alteration of the local conformation in the active site and thus crippling catalysis. Three mutations in the second category, including V151E and L155S in TM6, and S225I in TM7, may affect the protease activity through altered local conformation.

Perspective

Although the structure of PS1 can be modelled on the basis of considerable sequence homology with PSH, their functional aspects share few common themes. γ-secretase initially produces 48–49-residue amyloid-β, followed by successive cleavages of 3–4 amino acids at each step to produce a spectrum of peptides[8,9]. The main effect of mutations in PS1 is to increase the proportion of long, aggregation-prone amyloid-β peptides, exemplified by the Aβ42/Aβ40 ratio[8,9]. These functional aspects cannot be recapitulated in PSH. In this regard, the biochemical characterization of PSH has little implication on PS1 or γ-secretase.

A fundamental issue in the investigation of γ-secretase complex is how the four components are assembled. Our atomic model of PS1, together with available information from past studies, reveals insights into this important question. The TM4 of PS1 is thought to interact with the two TMs of PEN-2 (refs 12, 13). We took an *ab initio* structure prediction approach to generate atomic models of PEN-2 (Supplementary Fig. 13). One of the top scoring models was docked onto TM4 of the PS1 model, involving mostly hydrophobic interactions. Notably, the binding site for PEN-2 is on the opposite side of PS1-CTF, where APH-1 and nicastrin bind. We note that this docking model between PS1 and PEN-2 is highly speculative.

The oligomeric state of γ-secretase has been controversial[24], with evidence for a complex containing one molecule for each of the four components[24,37], or two molecules for PS1[38,39], or with an aggregated molecular mass of 900 kDa[40]. A recent electron microscopic study of γ-secretase revealed an accurate molecular mass of 230 kDa, consistent with the case of one molecule each for the four components[24]. However, as is the case for membrane protein, the oligomeric state and the solution behaviour of γ-secretase are likely to depend on the choice of solubilizing detergents. Given the high degree of sequence similarity between PSH and PS1, it is tempting to speculate that PS1 may also form a homo-tetramer. Alternatively, the observed tetrameric assembly of PSH could just be an artefact of *in vitro* manipulation and crystallization, or more likely a unique property of PSH. Regardless of these different scenarios, it is important to note that presenilin oligomerization is not required for γ-secretase activity[37].

Bacteria contain a family of membrane-embedded proteases preflagellin peptidase (PFP) and type-4 prepilin peptidase (TFPP). Unlike presenilin/SPP, PFP and TFPP cleave substrate proteins in aqueous environment. The structure of the GXGD protease FlaK, a PFP, was recently reported[26]. Similar to PSH, the two catalytic Asp residues are also uncoupled in FlaK. It should be noted, however, there is little sequence or structural similarity between PSH and FlaK (Supplementary Fig. 14). Most notably, PSH, but not FlaK, shares strong sequence homology with presenilin/SPP.

Although the focus of this manuscript is given to PS1, considerable sequence similarity exists between PSH and SPP (Supplementary Fig. 15). The structural information of PSH, together with sequence alignment, should allow accurate modelling of SPP structure and perhaps rationalization of functional data on SPP. For example, the identification of the C-terminal TMs as a minimalist functional core in SPP[41] is consistent with the observed domain organization of PSH. In fact, the degree of sequence similarity between PSH and human SPP is approximately 44%, similar to that between PSH and PS1 (Supplementary Fig. 15). Consequently, most of the structural features proposed for PS1 are likely to be conserved in human SPP as well. These features may include the overall TM organization and the locations of the active site aspartate residues. Intriguingly, electron microscopic analysis of human SPP reveals a tetrameric assembly[42] that is reminiscent of PSH.

In summary, structure of the aspartate intramembrane protease PSH, the first of its class, suggests important clues on the organization of nine TMs in PS1. Gratifyingly, our structural information is supported by a body of published experimental evidence. For one instance, TM1 of PS1 was shown to participate in the catalytic core structure[43]; in our structure, TM1 directly contacts TM7 and TM8 (Fig. 1b). In this regard, our study also serves as a framework for understanding the structure and mechanism of presenilin, γ-secretase and SPP.

METHODS SUMMARY

The recombinant PSH protein was overexpressed in *Escherichia coli* and purified to homogeneity by affinity chromatography and gel filtration. The PSH was crystallized by the hanging-drop vapour-diffusion method. Derivative crystals were obtained by soaking crystals for 20 min in mother liquor containing 2 mM K_2PtCl_4. Diffraction data were collected at the Shanghai Synchrotron Radiation Facility (SSRF) beamline BL17U and SPring-8 beamline BL41XU and processed with HKL2000 (ref. 44). The experimental phases were generated by Pt-MAD using SOLVE[45]. Multi-crystal averaging with all three available space groups (*C*222, $C222_1$, *P*2) combined with solvent flattening, histogram matching and non-crystallographic symmetry (NCS) averaging gave a map of sufficient quality for model building. An initial model was built manually using COOT[46]. Sequence docking was aided with the selenium sites in the Se-SAD anomalous difference Fourier map. The structure was refined with PHENIX[47]. The proteolytic activity of PSH was examined by an *in vitro* cleavage assay using Gurken as the substrate. Sequence alignment was carried out using ClustalW[48]. The human presenilin 1 (PS1) model was built according to the sequence alignment results with the Sculptor[49] from the PSH coordinates.

Full Methods and any associated references are available in the online version of the paper.

Received 6 September; accepted 16 November 2012.
Published online 19 December 2012.

1. Brown, M. S., Ye, J., Rawson, R. B. & Goldstein, J. L. Regulated intramembrane proteolysis: a control mechanism conserved from bacteria to humans. *Cell* **100,** 391–398 (2000).
2. Erez, E., Fass, D. & Bibi, E. How intramembrane proteases bury hydrolytic reactions in the membrane. *Nature* **459,** 371–378 (2009).
3. Urban, S. Making the cut: central roles of intramembrane proteolysis in pathogenic microorganisms. *Nature Rev. Microbiol.* **7,** 411–423 (2009).

4. Wolfe, M. S. *et al.* Two transmembrane aspartates in presenilin-1 required for presenilin endoproteolysis and γ-secretase activity. *Nature* **398,** 513–517 (1999).
5. De Strooper, B. *et al.* A presenilin-1-dependent γ-secretase-like protease mediates release of Notch intracellular domain. *Nature* **398,** 518–522 (1999).
6. Struhl, G. & Greenwald, I. Presenilin is required for activity and nuclear access of Notch in *Drosophila*. *Nature* **398,** 522–525 (1999).
7. De Strooper, B. *et al.* Deficiency of presenilin-1 inhibits the normal cleavage of amyloid precursor protein. *Nature* **391,** 387–390 (1998).
8. Selkoe, D. J. & Wolfe, M. S. Presenilin: running with scissors in the membrane. *Cell* **131,** 215–221 (2007).
9. De Strooper, B., Iwatsubo, T. & Wolfe, M. S. Presenilins and γ-secretase: structure, function, and role in Alzheimer disease. *Cold Spring Harb. Perspect. Med.* **2,** a006304 (2012).
10. Steiner, H. *et al.* Glycine 384 is required for presenilin-1 function and is conserved in bacterial polytopic aspartyl proteases. *Nature Cell Biol.* **2,** 848–851 (2000).
11. De Strooper, B. Aph-1, Pen-2, and Nicastrin with Presenilin generate an active γ-Secretase complex. *Neuron* **38,** 9–12 (2003).
12. Kim, S. H. & Sisodia, S. S. Evidence that the ''NF'' motif in transmembrane domain 4 of presenilin 1 is critical for binding with PEN-2. *J. Biol. Chem.* **280,** 41953–41966 (2005).
13. Watanabe, N. *et al.* Pen-2 is incorporated into the γ-secretase complex through binding to transmembrane domain 4 of presenilin 1. *J. Biol. Chem.* **280,** 41967–41975 (2005).
14. Takasugi, N. *et al.* The role of presenilin cofactors in the γ-secretase complex. *Nature* **422,** 438–441 (2003).
15. Fraering, P. C. *et al.* Detergent-dependent dissociation of active γ-secretase reveals an interaction between Pen-2 and PS1-NTF and offers a model for subunit organization within the complex. *Biochemistry* **43,** 323–333 (2004).
16. Chiang, P.-M., Fortna, R. R., Price, D. L., Li, T. & Wong, P. C. Specific domains in anterior pharynx-defective 1 determine its intramembrane interactions with nicastrin and presenilin. *Neurobiol. Aging* **33,** 277–285 (2012).
17. Wang, Y., Zhang, Y. & Ha, Y. Crystal structure of a rhomboid family intramembrane protease. *Nature* **444,** 179–180 (2006).
18. Wu, Z. *et al.* Structural analysis of a rhomboid family intramembrane protease reveals a gating mechanism for substrate entry. *Nature Struct. Mol. Biol.* **13,** 1084–1091 (2006).
19. Ben-Shem, A., Fass, D. & Bibi, E. Structural basis for intramembrane proteolysis by rhomboid serine proteases. *Proc. Natl Acad. Sci. USA* **104,** 462–466 (2007).
20. Lemieux, M. J., Fischer, S. J., Cherney, M. M., Bateman, K. S. & James, M. N. The crystal structure of the rhomboid peptidase from *Haemophilus influenzae* provides insight into intramembrane proteolysis. *Proc. Natl Acad. Sci. USA* **104,** 750–754 (2007).
21. Feng, L. *et al.* Structure of a site-2 protease family intramembrane metalloprotease. *Science* **318,** 1608–1612 (2007).
22. Lazarov, V. K. *et al.* Electron microscopic structure of purified, active γ-secretase reveals an aqueous intramembrane chamber and two pores. *Proc. Natl Acad. Sci. USA* **103,** 6889–6894 (2006).
23. Ogura, T. *et al.* Three-dimensional structure of the γ-secretase complex. *Biochem. Biophys. Res. Commun.* **343,** 525–534 (2006).
24. Osenkowski, P. *et al.* Cryoelectron microscopy structure of purified γ-secretase at 12 Å resolution. *J. Mol. Biol.* **385,** 642–652 (2009).
25. Sobhanifar, S. *et al.* Structural investigation of the C-terminal catalytic fragment of presenilin 1. *Proc. Natl Acad. Sci. USA* **107,** 9644–9649 (2010).
26. Hu, J., Xue, Y., Lee, S. & Ha, Y. The crystal structure of GXGD membrane protease FlaK. *Nature* **475,** 528–531 (2011).
27. Ponting, C. P. *et al.* Identification of a novel family of presenilin homologues. *Hum. Mol. Genet.* **11,** 1037–1044 (2002).
28. Torres-Arancivia, C. *et al.* Identification of an archaeal presenilin-like intramembrane protease. *PLoS ONE* **5,** e13072 (2010).
29. Kornilova, A. Y., Das, C. & Wolfe, M. S. Differential effects of inhibitors on the γ-secretase complex. Mechanistic implications. *J. Biol. Chem.* **278,** 16470–16473 (2003).
30. Sato, C., Morohashi, Y., Tomita, T. & Iwatsubo, T. Structure of the catalytic pore of γ-secretase probed by the accessibility of substituted cysteines. *J. Neurosci.* **26,** 12081–12088 (2006).
31. Tolia, A., Chavez-Gutierrez, L. & De Strooper, B. Contribution of presenilin transmembrane domains 6 and 7 to a water-containing cavity in the γ-secretase complex. *J. Biol. Chem.* **281,** 27633–27642 (2006).
32. Holm, L. & Sander, C. Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* **233,** 123–138 (1993).
33. Wu, S., Mehta, S. Q., Pichaud, F., Bellen, H. J. & Quiocho, F. A. Sec15 interacts with Rab11 via a novel domain and affects Rab11 localization *in vivo*. *Nature Struct. Mol. Biol.* **12,** 879–885 (2005).
34. Cooper, J. B., Khan, G., Taylor, G., Tickle, I. J. & Blundell, T. L. X-ray analyses of aspartic proteinases. II. Three-dimensional structure of the hexagonal crystal form of porcine pepsin at 2.3 Å resolution. *J. Mol. Biol.* **214,** 199–222 (1990).
35. Sato, C., Takagi, S., Tomita, T. & Iwatsubo, T. The C-terminal PAL motif and transmembrane domain 9 of presenilin 1 are involved in the formation of the catalytic pore of the γ-secretase. *J. Neurosci.* **28,** 6264–6271 (2008).
36. Kornilova, A. Y., Bihel, F., Das, C. & Wolfe, M. S. The initial substrate-binding site of γ-secretase is located on presenilin near the active site. *Proc. Natl Acad. Sci. USA* **102,** 3230–3235 (2005).
37. Sato, T. *et al.* Active γ-secretase complexes contain only one of each component. *J. Biol. Chem.* **282,** 33985–33993 (2007).
38. Schroeter, E. H. *et al.* A presenilin dimer at the core of the γ-secretase enzyme: insights from parallel analysis of Notch 1 and APP proteolysis. *Proc. Natl Acad. Sci. USA* **100,** 13075–13080 (2003).
39. Cervantes, S., Saura, C. A., Pomares, E., Gonzalez-Duarte, R. & Marfany, G. Functional implications of the presenilin dimerization. Reconstitution of γ-secretase activity by assembly of a catalytic site at the dimer interface of two catalytically inactive presenilins. *J. Biol. Chem.* **279,** 36519–36529 (2004).
40. Evin, G. *et al.* Transition-state analogue γ-secretase inhibitors stabilize a 900 kDa presenilin/nicastrin complex. *Biochemistry* **44,** 4332–4341 (2005).
41. Narayanan, S., Sato, T. & Wolfe, M. S. A C-terminal region of signal peptide peptidase defines a functional domain for intramembrane aspartic protease catalysis. *J. Biol. Chem.* **282,** 20172–20179 (2007).
42. Miyashita, H. *et al.* Three-dimensional structure of the signal peptide peptidase. *J. Biol. Chem.* **286,** 26188–26197 (2011).
43. Takagi, S., Tominaga, A., Sato, C., Tomita, T. & Iwatsubo, T. Participation of transmembrane domain 1 of presenilin 1 in the catalytic pore structure of the γ-secretase. *J. Neurosci.* **30,** 15943–159450 (2010).
44. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276,** 307–326 (1997).
45. Terwilliger, T. SOLVE and RESOLVE: automated structure solution and density modification. *J. Synchrotron Rad.* **11,** 49–52 (2004).
46. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60,** 2126–2132 (2004).
47. Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66,** 213–221 (2010).
48. Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22,** 4673–4680 (1994).
49. Bunkóczi, G. & Read, R. J. Improvement of molecular-replacement models with Sculptor. *Acta Crystallogr. D* **67,** 303–312 (2011).
50. DeLano, W. L. The PyMOL Molecular Graphics System. http://www.pymol.org (Schrödinger, 2002).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank J. He and Q. Wang at Shanghai Synchrotron Radiation Facility beamline BL17U and K. Hasegawa and T. Kumasaka at the SPring-8 beamline BL41XU for assistance. This work was supported by funds from the Ministry of Science and Technology (grant number 2009CB918801), and National Natural Science Foundation of China project 30888001.

Author Contributions X.L., S.D. and Y.S. designed all experiments. X.L., S.D., C.Y., X.G. and J.W. performed the experiments. All authors contributed to data analysis. X.L., S.D., X.G., J.W. and Y.S. contributed to manuscript preparation. Y.S. wrote the manuscript.

Author Information The atomic coordinates and structure factor files of PSH have been deposited in the Protein Data Bank with the accession codes 4HYG, 4HYD and 4HYC, respectively, for the space groups *C*222, *C*2221 and *P*2. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to Y.S. (shi-lab@tsinghua.edu.cn).

METHODS

Protein engineering. The complementary DNAs of 13 archaeal presenilin homologues were individually cloned into pET-21b (Novagen). The recombinant proteins were overexpressed in *E. coli* strain BL21(DE3). After first-round screening, the presenilin homologue from archaeon *Methanoculleus marisnigri* JR1 (with an N-terminal 8-His tag), named as PSH, became our focused target because of its reasonable yield. However, the solution behaviour of the wild-type, full-length PSH is poor, and the purified protein is highly prone to aggregation and precipitation. We embarked on a systematic journey designed to improve the solution behaviour of PSH. Sequence alignment of PSH with other sequence homologues in bacteria and eukaryotes revealed a large number of amino acids that are conserved in other homologues, but not in PSH. These amino acids were targeted for individual mutation (by conserved amino acids) and the variants were evaluated for solution behaviour and protease activity. Only those mutations that improve solution behaviour but have little impact on protease activity were chosen for double, triple, and quadruple mutations. Using this strategy, we generated more than 100 PSH variants. Many of these variants were crystallized. The variant that gave the best X-ray diffraction contains five mutations: D40N, E42S, A147E, V148P and A229V.

Protein expression and purification. The transformed BL21(DE3) cells were induced with 0.2 mM isopropyl-β-D-thiogalactopyranoside (IPTG) at an optical density of 1.5 at 600 nm. After growing for 16 h at 22 °C, the cells were collected, homogenized in a buffer containing 25 mM Tris-Cl, pH 8.0, and 150 mM NaCl, and lysed using a French Press with 2 passes at 15,000 p.s.i. Cell debris was removed by centrifugation. The supernatant was collected and applied to ultracentrifugation at 150,000*g* for 1 h. Membrane fraction was incubated with 2% (w/v) *n*-nonyl-β-D-maltopyranoside (NM, Anatrace) for 1 h at 4 °C. Following another ultracentrifugation step at 150,000*g* for 30 min, the supernatant was loaded onto Ni^{2+}-NTA affinity column (Qiagen). The resin was washed with 25 mM Tris-Cl, pH 8.0, 150 mM NaCl, 20 mM imidazole, 0.6% NM. The protein was eluted from the affinity resin by 25mM Tris-Cl, pH 8.0, 150 mM NaCl, 250 mM imidazole and 0.6% NM, and concentrated to about 20 mg ml^{-1} before further purification by gel filtration (Superdex-200, GE Healthcare). The buffer for gel filtration contained 25 mM Tris-Cl, pH 8.0, 150 mM NaCl, and 0.2% (w/v) *n*-nonyl-β-D-glucopyranoside (β-NG, Anatrace) plus 0.023% (w/v) *n*-dodecyl-N,N-dimethylamine-N-oxide (LDAO, Anatrace). All PSH variants were generated using two-step PCR and were subcloned, overexpressed and purified in the same way as wild-type protein. Selenomethionine (Se-Met)-labelled protein was purified similarly with the exception that 2 mM Tris-[2-carboxyethyl]-phosphine (TCEP) was included during purification.

The peak fraction of PSH from gel filtration was collected and incubated with the endoproteinase Glu-C from *Staphylococcus aureus* V8 (V8, Sigma-Aldrich) at 22 °C for 20 min, with a final concentration of 0.03 mg ml^{-1} for the V8 protease. After limited proteolysis, the resulting PSH protein was applied to gel filtration, which revealed two discrete fragments. Analysis by mass spectrometry revealed that an internal fragment in the extended loop connecting TM6 and TM7, residues 182–209, was removed by the V8 protease treatment. The resulting two-chain PSH protein was used for crystallization.

Crystallization. The concentration of protein for crystallization was approximately 6.5 mg ml^{-1}. Crystals were grown at 18 °C by the hanging-drop vapour-diffusion method. The digested protein gave rise to crystals of trapezoid plates. The crystals appeared 3 days in the well buffer containing 0.1 M glycine pH 3.6, 0.2 M $(NH_4)_2SO_4$, 20% (w/v) PEG500MME (Fluka), 6% (w/v) glycerol and 0.04% (w/v) Anapoe-$C_{12}E_8$ (Anatrace), and grew to full size in 3 weeks. These crystals are in three different space groups: *C*222, *C*222$_1$ and *P*2. Crystals of the *C*222 space group exhibited stronger diffraction than those of the other two space groups. Derivative crystals were obtained by soaking crystals for 20 min in mother liquor containing 2 mM K_2PtCl_4. Both native and heavy-atom-derived crystals were directly flash-frozen in a cold nitrogen stream at 100 K.

Data collection. The high-resolution native data was collected at Shanghai Synchrotron Radiation Facility (SSRF) beamline BL17U. Other diffraction data were collected at the SPring-8 beamline BL41XU. The platinum-soaked crystals are extremely sensitive to radiation damage, even at 100 K. To collect a complete, high-redundancy, 3-wavelength MAD data set, the size of the collimated X-ray beam was fine-tuned to maximize the number of exposures on a single crystal. For the same crystal, the three anomalous data sets (peak, inflection and high-energy remote) around the absorption edge of platinum were collected at three different positions, each for a different wavelength. All data were processed with the package HKL2000 (ref. 44) by routine procedure. The diffraction images from the severely anisotropic native crystal were collected at the SSRF beamline BL17U, and integrated with DENZO[44]. Before the .x files were input into SCALEPACK[44] for merging and scaling, the anisotropic ellipsoidal truncations on the .x files were performed with the special version of the ellipsoidal truncation program provided by the UCLA MBI – Diffraction Anisotropy Server[51]. The applied resolution limits along a*, b*, and c* directions are 3.32, 4.0, and 3.32 Å, respectively. No F/sigma cut-off was performed at any resolution shell within the ellipsoid. Selection of the resolution range was based on the criteria of F/sigma larger than 3.0 (see Supplementary Fig. 3a). The pruned raw data was then used for the final structural refinement. Further processing was carried out using programs from the CCP4 suites[52]. Data collection statistics are summarized in Supplementary Tables 1–3.

Structure determination. Six platinum atoms were determined using the program SHELXD[53]. Each PSH molecule binds at least one platinum atom through methionine residues. Four platinum atoms, one in each PSH molecule, display the same non-crystallographic symmetry (NCS) as that of the four PSH molecules in each asymmetric unit. The identified platinum sites were then refined and the initial phases were calculated by the program SOLVE[54] with the multi-wavelength anomalous dispersion (MAD) phasing method. The Figure-Of-Merit (FOM) right after SOLVE was measurable only up to 5 Å (black line in Supplementary Fig. 3b). The phases were refined and extended from 5 Å to 3.9 Å resolution by fourfold NCS (non-crystallographic symmetry) averaging, solvent flattening and histogram matching using DM from the *CCP4* program suite[52]. The resulting electron density map displays all nine transmembrane helices. A crude partial α-helices model was traced by hand, and the resulting model was then used for molecular replacement with the program PHASER[55] into the *C*222$_1$ and *P*2 crystal forms. Cross-crystal averaging with the three different crystal forms in Supplementary Table 1 combined with solvent flattening, histogram matching, and NCS averaging in DMMulti[56] yielded a map of sufficient quality for model building.

An initial model was built into the experimental electron density map using COOT[46]. Sequence docking was aided with the selenium sites in the anomalous difference Fourier map. The structure of the *C*222 space group was refined with PHENIX[57], using the following restraints: stereochemistry (including bond length, angle, torsion angle and chiral volume, etc.), NCS, secondary structure restraints (377 defined hydrogen bonds in α-helices), and experimental phases restraints. The refined model for one PSH molecule was used for molecular replacement into the *C*222$_1$ and *P*2 crystal forms, and the *C*222$_1$ and *P*2 structures were also refined using PHENIX[57] with the appropriate restraints.

We used selenium anomalous signals to help validate the atomic model. There are 12 Met residues in the sequence of PSH, not counting the initiation Met. Except one Met, all 11 other Met residues are located on six α-helices: TM1, TM2, TM3, TM6, TM7 and TM8. In the selenium anomalous difference Fourier map, all these 11 selenium sites were identified. However, there were no Met residues in TM4, TM5, and TM9. To confirm the sequence assignment for these three TMs, we generated three missense mutants of PSH: V114M on TM4, L133M on TM5, and L284M on TM9, individually purified these three proteins to homogeneity, and crystallized these variants. Single-wavelength anomalous diffraction data from these crystals were collected at the absorption edge of selenium. Data collection statistics were shown in Supplementary Table 3. The anomalous densities around V114M, L133M and L284M in the anomalous difference Fourier map were in complete agreement with the PSH atomic model (Supplementary Fig. 6a).

Protease activity assay. The proteolytic activity of PSH was examined by an *in vitro* cleavage assay. The MBP-fused transmembrane region of Gurken was used as the substrate[28]. The assay was performed at 37 °C for 8 h in a buffer containing PBS, 0.02% (w/v) DDM and 20 mM citrate pH 5.3. The concentrations of the substrate protein and protease were approximately 0.5 mg ml^{-1}. N-[[[(2R,3S)-3-[[(1,1-dimethylethoxy)carbonyl]amino]-2-hydroxy-4-phenylbutyl](phenylmethyl)amino]carbonyl]-L-leucyl-L-valine methyl ester WPE-III-31-C (abbreviated as III-31-C, Sigma-Aldrich), a γ-secretase inhibitor[29] was dissolved in DMSO and used for inhibition assay. The reaction was stopped by SDS sample buffer and the cleavage products were analysed by SDS–PAGE and Coomassie staining.

Structural modelling of human presenilin 1 (PS1). Presenilin 1 sequences from *Homo sapiens*, *Xenopus laevis* and *Drosophila melanogaster* were aligned to their archaeal homologue (PSH) based on the multiple sequence alignment algorithm implemented in the ClustalW[48]. The human presenilin 1 (PS1) model was built according to the sequence alignment results with the Sculptor[49] from the PSH coordinates. The missing side chains were built manually in COOT[46], and the missing loops were built with the program SLOOP[58] by using fragments from the Richardson's Top500 library of structures to fill the gaps. The model geometry was regularized with the program phenix.pdbtools –geometry_regularization[47].

Structure prediction of PEN-2. The three-dimensional structure of PEN-2 was predicted by *ab initio* method because there is no published structure of similar sequences. In this method, amino acid sequences, fragment structure templates, and membrane environment were used as input. The three-residue fragments and nine-residue fragments were generated from the ROBETTA fragment server (http://robetta.bakerlab.org/fragmentsubmit.jsp). The membrane environment was defined from the predicted membrane normal and centre vectors from predicted trans-membrane region using the method OCTOPUS[59]. The membrane

ab initio modelling protocol[60] from Rosetta 3.3 was implemented to take advantage of the above information to generate 2000 models. The models with high scores and good topologies were selected as candidate structures.

51. Strong, M. *et al.* Toward the structural genomics of complexes: crystal structure of a PE/PPE protein complex from *Mycobacterium tuberculosis*. *Proc. Natl Acad. Sci. USA* **103,** 8060–8065 (2006).
52. Collaborative Computational Project, number 4. The *CCP*4 suite: programs for protein crystallography. *Acta Crystallogr. D* **50,** 760–763 (1994).
53. Schneider, T. R. & Sheldrick, G. M. Substructure solution with SHELXD. *Acta Crystallogr. D* **58,** 1772–1779 (2002).
54. Terwilliger, T. C. & Berendzen, J. Automated structure solution for MIR and MAD. *Acta Crystallogr. D* **55,** 849–861 (1999).
55. McCoy, A. J. *et al.* Phaser crystallographic software. *J. Appl. Crystallogr.* **40,** 658–674 (2007).
56. Cowtan, K. 'dm': an automated procedure for phase improvement by density modification. *Joint CCP4 ESF-EACBM Newslett. Prot. Crystallogr.* **31,** 34–38 (1994).
57. Adams, P. D. *et al.* PHENIX: building new software for automated crystallographic structure determination. *Acta Crystallogr. D* **58,** 1948–1954 (2002).
58. Cowtan, K. Completion of autobuilt protein models using a database of protein fragments. *Acta Crystallogr. D* **68,** 328–335 (2012).
59. Viklund, H. & Elofsson, A. OCTOPUS: improving topology prediction by two-track ANN-based preference scores and an extended topological grammar. *Bioinformatics* **24,** 1662–1668 (2008).
60. Barth, P., Wallner, B. & Baker, D. Prediction of membrane protein structures with complex topologies using limited constraints. *Proc. Natl Acad. Sci. USA* **106,** 1409–1414 (2009).

LETTER

doi:10.1038/nature11717

A vast, thin plane of corotating dwarf galaxies orbiting the Andromeda galaxy

Rodrigo A. Ibata[1], Geraint F. Lewis[2], Anthony R. Conn[3], Michael J. Irwin[4], Alan W. McConnachie[5], Scott C. Chapman[6], Michelle L. Collins[7], Mark Fardal[8], Annette M. N. Ferguson[9], Neil G. Ibata[10], A. Dougal Mackey[11], Nicolas F. Martin[1,7], Julio Navarro[12], R. Michael Rich[13], David Valls-Gabaud[14] & Lawrence M. Widrow[15]

Dwarf satellite galaxies are thought to be the remnants of the population of primordial structures that coalesced to form giant galaxies like the Milky Way[1]. It has previously been suspected[2] that dwarf galaxies may not be isotropically distributed around our Galaxy, because several are correlated with streams of H I emission, and may form coplanar groups[3]. These suspicions are supported by recent analyses[4–7]. It has been claimed[7] that the apparently planar distribution of satellites is not predicted within standard cosmology[8], and cannot simply represent a memory of past coherent accretion. However, other studies dispute this conclusion[9–11]. Here we report the existence of a planar subgroup of satellites in the Andromeda galaxy (M 31), comprising about half of the population. The structure is at least 400 kiloparsecs in diameter, but also extremely thin, with a perpendicular scatter of less than 14.1 kiloparsecs. Radial velocity measurements[12–15] reveal that the satellites in this structure have the same sense of rotation about their host. This shows conclusively that substantial numbers of dwarf satellite galaxies share the same dynamical orbital properties and direction of angular momentum. Intriguingly, the plane we identify is approximately aligned with the pole of the Milky Way's disk and with the vector between the Milky Way and Andromeda.

We undertook the Pan-Andromeda Archaeological Survey[16] (PAndAS) to obtain a large-scale panorama of the halo of the Andromeda galaxy (M 31), a view that is not available to us for the Milky Way. This Canada–France–Hawaii Telescope survey imaged about 400 square degrees around M 31, which is the only giant galaxy in the Local Group besides our Milky Way. Stellar objects are detected out to a projected distance of about 150 kiloparsecs (kpc) from M 31, and about 50 kpc from M 33, the most massive satellite of M 31. The data reveal a substantial population of dwarf spheroidal galaxies that accompany Andromeda[17].

The distances to the dwarf galaxies can be estimated by measuring the magnitude of the tip of the red-giant branch[18]. Improving on earlier methods, we have developed a Bayesian approach that yields the probability distribution function for the distance to each individual satellite[19]. In this way we now have access to homogeneous distance measurements (typical uncertainties 20–50 kpc) to the 27 dwarf galaxies (filled circles in Fig. 1) visible within the PAndAS survey area[20] that lie beyond the central 2.5°.

In Fig. 2 these distance measurements are used to calculate the sky positions of the homogeneous sample of 27 dwarf galaxies as they would appear from the centre of the Andromeda galaxy. Visually, there appears to be a correlation close to a particular great circle (red line): this suggests that there is a plane, centred on M 31, around which a subsample of the satellites have very little scatter. This is confirmed by the Monte Carlo analysis presented in the Supplementary Information, where we show that the probability of the alignment of the subsample of $n_{\rm sub} = 15$ satellites marked red in Figs 1 and 2 occurring at random is 0.13% (see Supplementary Fig. 1).

Following this discovery, we sought to investigate whether the subsample displayed any kinematic coherence. The radial velocity of each satellite is shown in Fig. 3, corrected for the bulk motion of the Andromeda system towards us: what is immediately striking is that 13 out of the 15 satellites possess coherent rotational motion, such that the southern satellites are approaching us with respect to M 31, while the northern satellites recede away from us with respect to their host galaxy.

The probability that 13 or more out of 15 objects should share the same sense of rotation is 1.4% (allowing for right-handed or left-handed rotation). Thus the kinematic information confirms the spatial correlation initially suspected from a visual inspection of Fig. 2. The total significance of the planar structure is approximately 99.998%.

Thus we conclude that we have detected, with very high confidence, a coherent planar structure of 13 satellites with a root-mean-square thickness of 12.6 ± 0.6 kpc (<14.1 kpc at 99% confidence), that corotate around M 31 with a (right-handed) axis of rotation that points approximately east. The three-dimensional configuration can be assessed visually in Fig. 3. The extent of the structure is gigantic, over 400 kpc along the line of sight and nearly 300 kpc north-to-south. Indeed, since Andromeda XIV and Cassiopeia II lie at the southern and northern limits of the PAndAS survey, respectively, it is quite probable that additional (faint) satellites belonging to this structure are waiting to be found just outside the PAndAS footprint. Although huge in extent, the structure appears to be lopsided, with most of the satellites populating the side of the halo of M 31 that is nearest to the Milky Way. The completeness analysis we have undertaken shows that this configuration is not due to a lowered detection sensitivity at large distance, but reflects a true paucity of satellites in the more distant halo hemisphere[21].

The existence of this structure had been hinted at in earlier work[22], thanks to the planar alignment we report on here being viewed nearly edge-on, although the grouping could not be shown to be statistically significant from the information available at that time. Other previously claimed alignments do not match the present plane, although they share some of the member galaxies: the most significant alignment of ref. 23 has a pole 45° away from that found here, and the (tentative) poles of the configuration in ref. 4 are 23.4° away, whereas those of a later contribution[6] have poles 34.1° and 25.4° distant from

[1]Observatoire Astronomique de Strasbourg, 11 rue de l'Université, F-67000 Strasbourg, France. [2]Sydney Institute for Astronomy, School of Physics, A28, The University of Sydney, New South Wales 2006, Australia. [3]Department of Physics and Astronomy, Macquarie University, New South Wales 2109, Australia. [4]Institute of Astronomy, University of Cambridge, Madingley Road, Cambridge CB3 0HA, UK. [5]NRC Herzberg Institute of Astrophysics, 5071 West Saanich Road, Victoria, British Columbia, V9E 2E7, Canada. [6]Department of Physics and Atmospheric Science, Dalhousie University, 6310 Coburg Road, Halifax, Nova Scotia, B3H 4R2, Canada. [7]Max-Planck-Institut für Astronomie, Königstuhl 17, 69117 Heidelberg, Germany. [8]University of Massachusetts, Department of Astronomy, LGRT 619-E, 710 North Pleasant Street, Amherst, Massachusetts 01003-9305, USA. [9]Institute for Astronomy, University of Edinburgh, Royal Observatory, Blackford Hill, Edinburgh EH9 3HJ, UK. [10]Lycée International des Pontonniers, 1 rue des Pontonniers, F-67000 Strasbourg, France. [11]The Australian National University, Mount Stromlo Observatory, Cotter Road, Weston Creek, Australian Capital Province 2611, Australia. [12]Department of Physics and Astronomy, University of Victoria, 3800 Finnerty Road, Victoria, British Columbia, V8P 5C2, Canada. [13]Department of Physics and Astronomy, University of California, Los Angeles, PAB, 430 Portola Plaza, Los Angeles, California 90095-1547, USA. [14]LERMA, UMR CNRS 8112, Observatoire de Paris, 61 Avenue de l'Observatoire, 75014 Paris, France. [15]Department of Physics, Engineering Physics, and Astronomy, Queen's University, Kingston, Ontario, K7L 3N, Canada.

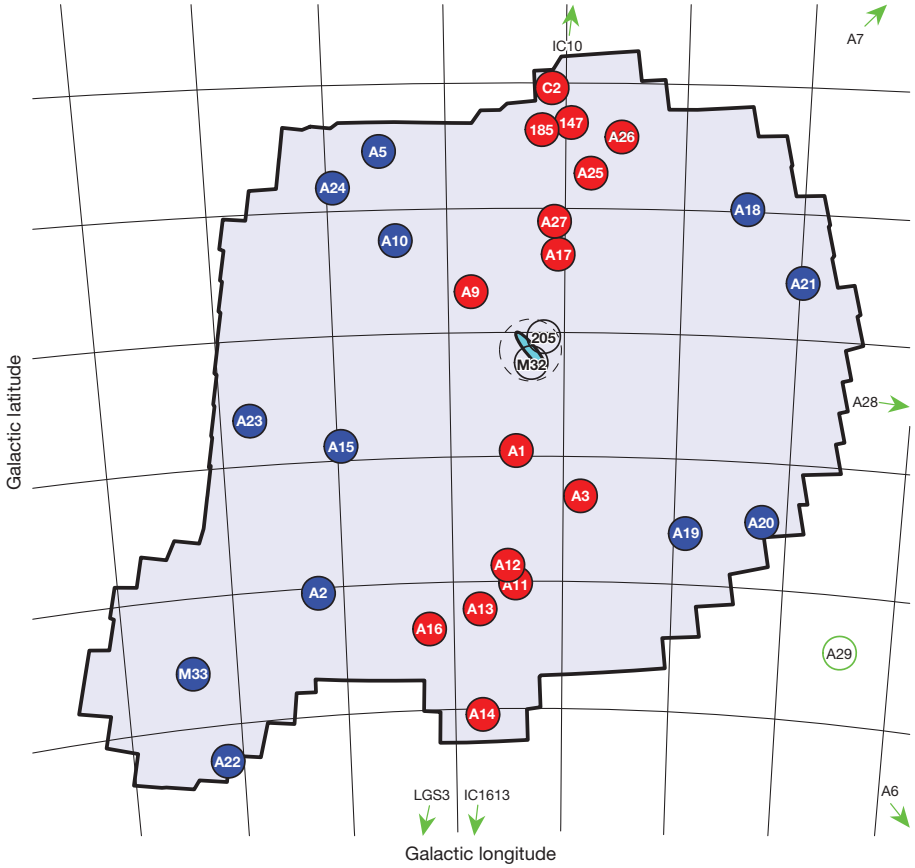


Figure 1 | Map of the Andromeda satellite system. The homogeneous PAndAS survey (irregular polygon) provides the source catalogue for the detection and distance measurements of the 27 satellite galaxies[20] (filled circles) used in this study. Near M 31 (blue ellipse), the high background hampers the detection of new satellites and precludes reliable distance measurements for M 32 and NGC 205 (labelled black open circles); we therefore exclude the region inside 2.5° (dashed circle) from the analysis. The seven satellites known outside the PandAS area (green circles and arrows) constitute a heterogeneous sample, discovered in various surveys with non-uniform spatial coverage, and their distances are not measured in the same homogeneous way. A reliable spatial analysis requires a data set with homogeneous selection criteria, so we do not include these objects in the sample either. The analysis shows that the satellites marked red are confined to a highly planar structure. We note that this structure is approximately perpendicular to lines of constant Galactic latitude, so it is therefore aligned approximately perpendicular to the Milky Way's disk (the grid squares are $4^\circ \times 4^\circ$).

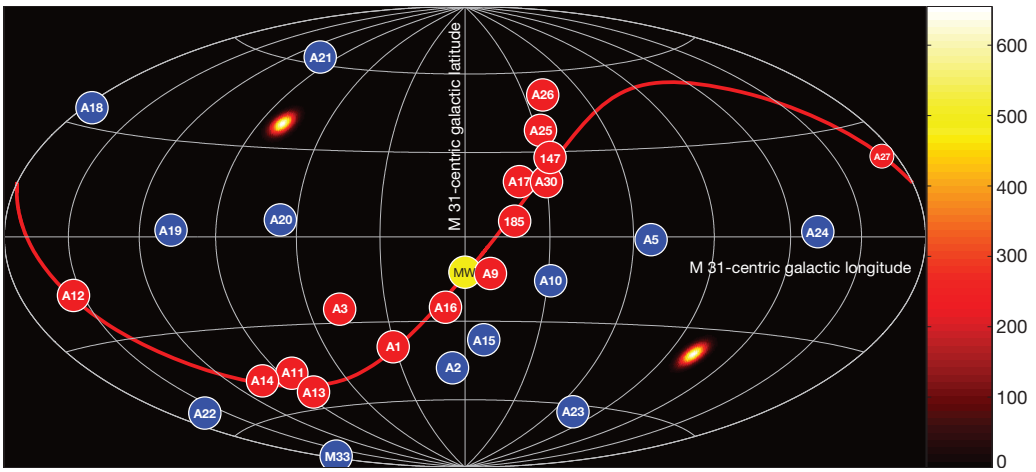
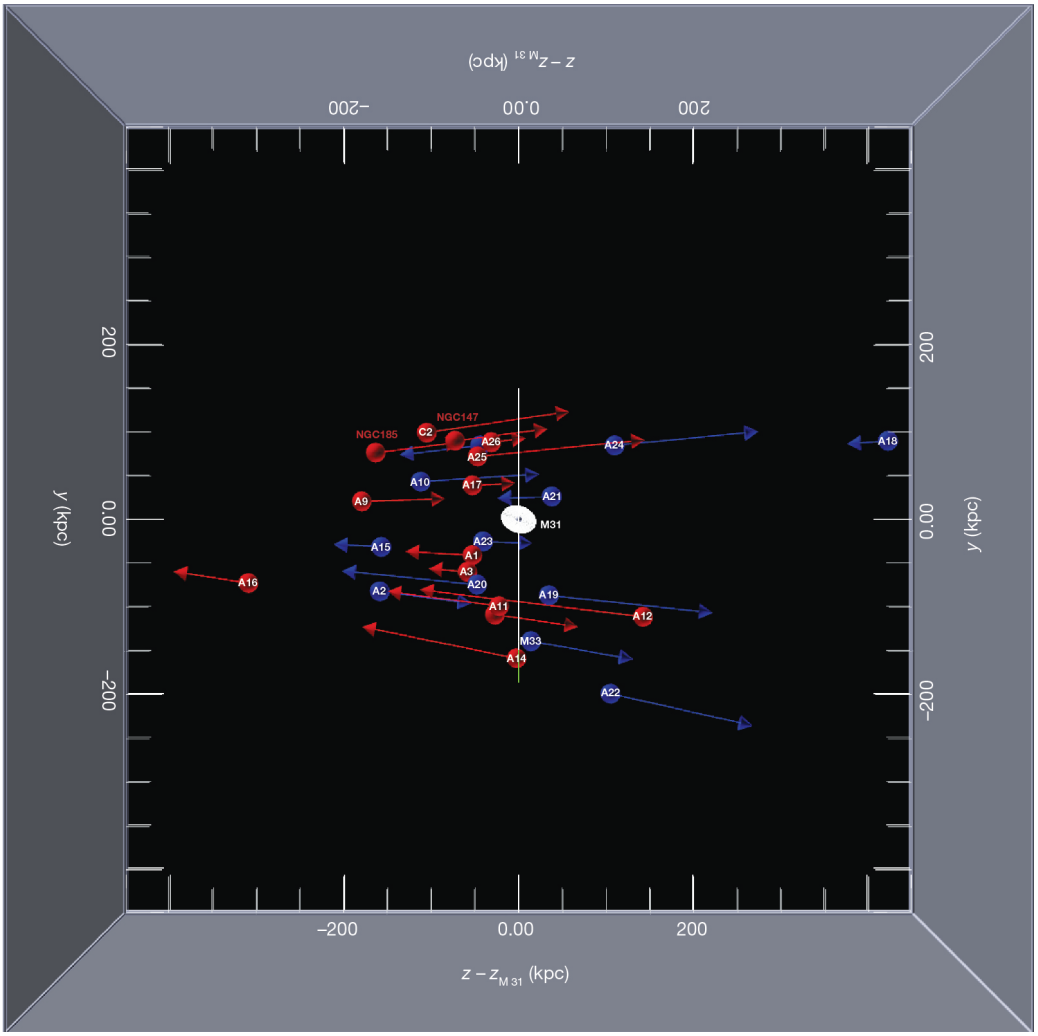


Figure 2 | Satellite galaxy positions as viewed from Andromeda. The Aitoff–Hammer projection shows the sample of 27 satellites[20] (filled circles from Fig. 1) as they would be seen from the centre of the Andromeda galaxy. In these coordinates the disk of Andromeda lies along the equator. 'M 31-centric galactic latitude' means what a fictitous observer in the M31 galaxy would call 'galactic latitude'. The background image represents the probability density function of the poles derived from 10^5 iterations of resampling the 27 satellites from their distance probability density functions, and finding the plane of lowest root mean square from a subsample of 15 (the colour scale on the right shows the relative probability of the poles, and is dimensionless). A clear narrow peak at ($l_{\mathrm{M31}} = 100.9^\circ \pm 0.9^\circ$, $b_{\mathrm{M31}} = -38.2^\circ \pm 1.4^\circ$) highlights the small uncertainty in the best-fit plane. The solid red line, which passes within less than 1° of the position of the Milky Way (yellow circle labelled 'MW'), represents the plane corresponding to this best pole location.



Click figure to activate

Figure 3 | Three-dimensional view (online only in the PDF version) or two-dimensional screenshot (in the print version) of the planar, rotating structure. The coordinate system is such that the z direction is parallel to the vector pointing from the Milky Way to M 31, x increases eastwards and y northwards. Only the radial component of the velocity of each satellite is measured, and these velocities are shown as vectors pointing either towards or away from the Milky Way. As in Figs 1 and 2, red spheres mark the planar satellites, and blue spheres represent the 'normal' population. The coherent kinematic behaviour of the spatially very thin structure (red) is clearly apparent viewed from the y–z plane. With the exception of And XIII and And XXVII, the satellites in the planar structure that lie to the north of M 31 recede from us, whereas those to the South approach us; this property strongly suggests rotation. Our velocity measurements[15] (supplemented by values from the literature[14]), have very small uncertainties, typically $<5\,\mathrm{km\,s^{-1}}$. The irregular green polygon (visible only in the x–y plane of the three-dimensional online version) shows the PAndAS survey area, the white circle (visible only in the x–y plane of the three-dimensional online version) indicates a projected radius of 150 kpc at the distance of M 31, and the white arrow (visible only in the three-dimensional online version) marks a velocity scale of $100\,\mathrm{km\,s^{-1}}$. (And XXVII is not shown in this diagram because its most likely distance is 476 kpc behind M 31). This figure is three-dimensionally interactive in the online version (allowing the reader to change the magnification and viewing angle), and was constructed with the S2PLOT programming library[26].

ours. Without the increased sample size, reliable three-dimensional positions and radial velocities, and most importantly, a spatially unbiased selection function resulting from the homogeneous panoramic coverage of PAndAS, the nature, properties and conclusive statistical significance of the present structure could not be inferred.

The present detection proves that in some giant galaxies, a significant fraction of the population of dwarf satellite galaxies, in this case around 50% (13 out of 27 over the homogeneously surveyed PAndAS area), are aligned in coherent planar structures, sharing the same direction of angular momentum. The Milky Way is the only other giant galaxy where we have access to high-quality three-dimensional positional data, and the existence of a similar structure around our Galaxy is strongly suggested by current data[2,5,7]. The implications for the origin and dynamical history of dwarf galaxies are profound. It also has a strong bearing on the analyses of dark matter in these darkest of galaxies, because one cannot now justifiably assume such objects to have evolved in dynamical isolation.

Intriguingly, the Milky Way lies within 1° of the plane reported here, the pole of the plane and the pole of the Milky Way's disk are approximately perpendicular (81°), and furthermore this plane is approximately perpendicular to the plane of satellites that has been proposed to surround the Galaxy (given that its pole points approximately towards Andromeda[7] within the uncertainties). Although these alignments may be chance occurrences, it is nevertheless essential information about the structure of the nearby Universe that must be taken into account in future simulations aimed at modelling the dynamical formation history of the Local Group.

The formation of this structure around M 31 poses a puzzle. For discussion, we envisage two broad classes of possible explanations: accretion or *in situ* formation. In either type of model, the small scatter

of the satellites out of the plane is difficult to explain, even though the orbital timescales for the satellites are long (around 5 Gyr for satellites at 150 kpc). All the galaxies in the plane are known to have old, evolved, stellar populations, and so *in situ* formation would additionally imply that the structure is ancient.

In an accretion scenario, the dynamical coherence points to an origin in a single accretion of a group of dwarf galaxies. However, the spatial extent of the progenitor group would have to be broadly equal to or smaller than the current plane thickness ($<$14.1 kpc), yet no such groups are known. Interpreting the coherent rotation as a result of our viewing perspective[24] requires a bulk tangential velocity for the in-falling group of the order of 1,000 km s^{-1}, which seems unphysically high. A further possibility is that we are witnessing accretion along filamentary structures that are fortuitously aligned. *In situ* formation may be possible if the planar satellite galaxies formed like tidal-dwarf galaxies in an ancient gas-rich galaxy merger[7], but then the dwarf galaxies should be essentially devoid of dark matter. If the planar M 31 dwarfs are dynamically relaxed, the absence of dark matter would be greatly at odds with inferences from detailed observations[25] of Milky Way satellites, assuming the standard theory of gravity. An alternative possibility is that gas was accreted preferentially onto dark matter sub-halos that were already orbiting in this particular plane, but then the origin of the plane of sub-haloes would still require explanation. We conclude that it remains to be seen whether galaxy formation models within the current cosmological framework can explain the existence of this vast, thin, rotating structure of dwarf galaxies within the halo of our nearest giant galactic neighbour.

Received 18 September; accepted 23 October 2012.

1. Klypin, A., Kravtsov, A. V., Valenzuela, O. & Prada, F. Where are the missing galactic satellites? *Astrophys. J.* **522,** 82–92 (1999).
2. Lynden-Bell, D. Dwarf galaxies and globular clusters in high velocity hydrogen streams. *Mon. Not. R. Astron. Soc.* **174,** 695–710 (1976).
3. Lynden-Bell, D. & Lynden-Bell, R. M. Ghostly streams from the formation of the Galaxy's halo. *Mon. Not. R. Astron. Soc.* **275,** 429–442 (1995).
4. Metz, M., Kroupa, P. & Jerjen, H. The spatial distribution of the Milky Way and Andromeda satellite galaxies. *Mon. Not. R. Astron. Soc.* **374,** 1125–1145 (2007).
5. Metz, M., Kroupa, P. & Libeskind, N. I. The orbital poles of Milky Way satellite galaxies: a rotationally supported disk of satellites. *Astrophys. J.* **680,** 287–294 (2008).
6. Metz, M., Kroupa, P. & Jerjen, H. Discs of satellites: the new dwarf spheroidals. *Mon. Not. R. Astron. Soc.* **394,** 2223–2228 (2009).
7. Pawlowski, M. S., Pflamm-Altenburg, J. & Kroupa, P. The VPOS: a vast polar structure of satellite galaxies, globular clusters and streams around the Milky Way. *Mon. Not. R. Astron. Soc.* **423,** 1109–1126 (2012).
8. Komatsu, E. *et al.* Seven-year Wilkinson Microwave Anisotropy Probe (WMAP) observations: cosmological interpretation. *Astrophys. J.* **192** (Suppl.), 18 (2011).
9. Zentner, A. R., Kravtsov, A. V., Gnedin, O. Y. & Klypin, A. A. The anisotropic distribution of galactic satellites. *Astrophys. J.* **629,** 219–232 (2005).
10. Lovell, M. R., Eke, V. R., Frenk, C. S. & Jenkins, A. The link between galactic satellite orbits and subhalo accretion. *Mon. Not. R. Astron. Soc.* **413,** 3013–3021 (2011).
11. Wang, J., Frenk, C. S., Navarro, J. F., Gao, L. & Sawala, T. The missing massive satellites of the Milky Way. *Mon. Not. R. Astron. Soc.* **424,** 2715–2721 (2012).
12. Collins, M. L. M. *et al.* The scatter about the 'Universal' dwarf spheroidal mass profile: a kinematic study of the M31 satellites And V and And VI. *Mon. Not. R. Astron. Soc.* **417,** 1170–1182 (2011).
13. Tollerud, E. J. *et al.* The SPLASH survey: spectroscopy of 15 M31 dwarf spheroidal satellite galaxies. *Astrophys. J.* **752,** 45 (2012).
14. McConnachie, A. W. The observed properties of dwarf galaxies in and around the Local Group. *Astron. J.* **144,** 4 (2012).
15. Collins, M. L. *et al.* The non-universal dSpH mass profile? A kinematic study of the Andromeda dwarf spheroidal system. *Mon. Not. R. Astron. Soc.* (submitted).
16. McConnachie, A. W. *et al.* The remnants of galaxy formation from a panoramic survey of the region around M 31. *Nature* **461,** 66–69 (2009).
17. Richardson, J. C. *et al.* PAndAS' progeny: extending the M31 dwarf galaxy cabal. *Astrophys. J.* **732,** 76 (2011).
18. Lee, M. G., Freedman, W. L. & Madore, B. F. The tip of the red giant branch as a distance indicator for resolved galaxies. *Astrophys. J.* **417,** 553–559 (1993).
19. Conn, A. R. *et al.* A Bayesian approach to locating the red giant branch tip magnitude. I. *Astrophys. J.* **740,** 69 (2011).
20. Conn, A. R. *et al.* A Bayesian approach to locating the red giant branch tip magnitude. II. Distances to the satellites of M31. *Astrophys. J.* **758,** 11 (2012). 1209.4952.
21. McConnachie, A. W. & Irwin, M. J. The satellite distribution of M31. *Mon. Not. R. Astron. Soc.* **365,** 902–914 (2006).
22. Majewski, S. R. *et al.* Discovery of Andromeda XIV: a dwarf spheroidal dynamical rogue in the Local Group? *Astrophys. J. Lett.* **670,** L9–L12 (2007).
23. Koch, A. & Grebel, E. K. The anisotropic distribution of M31 satellite galaxies: a polar great plane of early-type companions. *Astron. J.* **131,** 1405–1415 (2006).
24. van der Marel, R. P. & Guhathakurta, P. M31 transverse velocity and local group mass from satellite kinematics. *Astrophys. J.* **678,** 187–199 (2008).
25. Walker, M. G. Dark matter in the Milky Way's dwarf spheroidal satellites. Preprint at http://arxiv.org/abs/1205.0311 (2012).
26. Barnes, D. G., Fluke, C. J., Bourke, P. D. & Parry, O. T. An advanced, three-dimensional plotting library for astronomy. *Publ. Astron. Soc. Aust.* **23,** 82–93 (2006).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank the staff of the Canada-France-Hawaii Telescope for taking the PAndAS data, and for their continued support throughout the project. We thank one of our referees, B. Tully, for pointing out that IC 1613 could also be associated to the planar structure. R.A.I. and D.V.G. gratefully acknowledge support from the Agence Nationale de la Recherche though the grant POMMME, and would like to thank B. Famaey for discussions. G.F.L. thanks the Australian Research Council for support through his Future Fellowship and Discovery Project. This work is based on observations obtained with MegaPrime/MegaCam, a joint project of CFHT and CEA/DAPNIA, at the Canada–France–Hawaii Telescope, which is operated by the National Research Council (NRC) of Canada, the Institut National des Sciences de l'Univers of the Centre National de la Recherche Scientifique (CNRS) of France, and the University of Hawaii. Some of the data presented here were obtained at the W.M. Keck Observatory, which is operated as a scientific partnership among the California Institute of Technology, the University of California and the National Aeronautics and Space Administration. The Observatory was made possible by the generous financial support of the W.M. Keck Foundation.

Author Contributions All authors assisted in the development and writing of the paper. In addition, the structural and kinematic properties of the dwarf population, and the significance of the Andromeda plane were determined by R.A.I., G.F.L. and A.R.C., based on distances determined by the same group (as part of the PhD research of A.R.C.). In addition, A.W.M. is the Principal Investigator of PAndAS; M.J.I. and R.A.I. led the data processing effort; R.A.I. was the Principal Investigator of an earlier CFHT MegaPrime/MegaCam survey, which PAndAS builds on (which included S.C.C., A.M.N.F., M.J.I., G.F.L., N.F.M. and A.W.M.). R.M.R. is Principal Investigator of the spectroscopic follow-up with the Keck Telescope. M.L.C. and S.C.C. led the analysis of the kinematic determination of the dwarf population, and N.F.M. led the detection of the dwarf population from PAndAS data. N.G.I. performed the initial analysis of the satellite kinematics.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to R.A.I. (rodrigo.ibata@astro.unistra.fr).

LETTER

doi:10.1038/nature11734

Giant magnetized outflows from the centre of the Milky Way

Ettore Carretti[1], Roland M. Crocker[2,3], Lister Staveley-Smith[4,5], Marijke Haverkorn[6,7], Cormac Purcell[8], B. M. Gaensler[8], Gianni Bernardi[9], Michael J. Kesteven[10] & Sergio Poppi[11]

The nucleus of the Milky Way is known to harbour regions of intense star formation activity as well as a supermassive black hole[1]. Recent observations have revealed regions of γ-ray emission reaching far above and below the Galactic Centre (relative to the Galactic plane), the so-called 'Fermi bubbles'[2]. It is uncertain whether these were generated by nuclear star formation or by quasar-like outbursts of the central black hole[3–6] and no information on the structures' magnetic field has been reported. Here we report observations of two giant, linearly polarized radio lobes, containing three ridge-like substructures, emanating from the Galactic Centre. The lobes each extend about 60 degrees in the Galactic bulge, closely corresponding to the Fermi bubbles, and are permeated by strong magnetic fields of up to 15 microgauss. We conclude that the radio lobes originate in a biconical, star-formation-driven (rather than black-hole-driven) outflow from the Galaxy's central 200 parsecs that transports a huge amount of magnetic energy, about 10^{55} ergs, into the Galactic halo. The ridges wind around this outflow and, we suggest, constitute a 'phonographic' record of nuclear star formation activity over at least ten million years.

We use the images of the recently concluded S-band Polarization All Sky Survey (S-PASS) that has mapped the polarized radio emission of the entire southern sky. The survey used the Parkes Radio Telescope at a frequency of 2,307 MHz, with 184 MHz bandwidth, and 9′ angular resolution[7].

The lobes we report here exhibit diffuse polarized emission (Fig. 1), an integrated total intensity flux of 21 kJy, and a high polarization fraction of 25%. They trace the Fermi bubbles excepting the top western (that is, right) corners where they extend beyond the region covered by the γ-ray emission structure. Depolarization by H II regions establishes that the lobes are almost certainly associated with the Galactic Centre (Fig. 2 and Supplementary Information), implying that their height is ~8 kpc. Archival data of WMAP[8] reveal the same structures at a microwave frequency of 23 GHz (Fig. 3). The 2.3–23 GHz spectral index α (with flux density S at frequency ν modelled as $S_\nu \propto \nu^{\alpha}$) of linearly polarized emission interior to the lobes spans the range −1.0 to −1.2, generally steepening with projected distance from the Galactic plane (see Supplementary Information). Along with the high polarization fraction, this phenomenology indicates that the lobes are due to cosmic-ray electrons, transported from the plane, synchrotron-radiating in a partly ordered magnetic field.

Three distinct emission ridges that all curve towards Galactic west with increasing Galactic latitude are visible within the lobes (Fig. 1); two other substructures proceeding roughly northwest and southwest from around the Galactic Centre hint at limb brightening in the biconical base of the lobes. These substructures all have counterparts in WMAP polarization maps (Fig. 3), and one of them[9], already known from radio continuum data as the Galactic Centre spur[10], appears to connect back to the Galactic Centre; we label the other substructures the northern and southern ridges. The ridges' magnetic field directions (Fig. 3) curve, following their structures. The Galactic Centre spur and southern ridges also seem to have GeV γ-ray counterparts (Fig. 2; also compare ref. 3). The two limb brightening spurs at the biconical lobe base are also visible in the WMAP map, where they appear to connect back to the Galactic Centre area. A possible third spur develops northeast from the Galactic Centre. These limb brightening spurs are also obvious in the Stokes U map as an X-shaped structure centred at the Galactic Centre (Supplementary Fig. 3).

Such coincident, non-thermal radio, microwave and γ-ray emission indicates the presence of a non-thermal electron population covering at least the energy range 1–100 GeV (Fig. 4) that is simultaneously synchrotron-radiating at radio and microwave frequencies and upscattering ambient radiation into γ-rays by the inverse Compton process. The widths of the ridges are remarkably constant at ~300 pc over their lengths. The ridges have polarization fractions of 25–31% (see Supplementary Information), similar to the average over the lobes. Given this emission and the stated polarization fractions, we infer magnetic field intensities of 6–12 μG for the lobes and 13–15 μG for the ridges (see Figs 2 and 3, and Supplementary Information).

An important question about the Fermi bubbles is whether they are ultimately powered by star formation or by activity of the Galaxy's central, supermassive black hole. Despite their very large extent, the γ-ray bubbles and the X-shaped polarized microwave and X-ray structures tracing their limb-brightened base[11] have a narrow waist of only 100–200 pc diameter at the Galactic Centre. This matches the extent of the star-forming molecular gas ring (of $\sim 3 \times 10^7$ solar masses) recently demonstrated to occupy the region[12]. With 5–10% of the Galaxy's molecular gas content[1], star-formation activity in this 'central molecular zone' is intense, accelerating a distinct cosmic ray population[13,14] and driving an outflow[11,15] of hot, thermal plasma, cosmic rays and 'frozen-in' magnetic field lines[6,14,16].

One consequence of the region's outflow is that the cosmic ray electrons accelerated there (dominantly energized by supernovae) are advected away before they lose much energy radiatively *in situ*[14,16,17]. This is revealed by the fact that the radio continuum flux on scales up to 800 pc around the Galactic Centre is in anomalous deficit with respect to the expectation afforded by the empirical far-infrared/radio continuum correlation[18]. The total 2.3 GHz radio continuum flux from the lobes of ~21 kJy, however, saturates this correlation as normalized to the 60 μm flux (2 MJy) of the inner ~160 pc diameter region (ref. 19). Together with the morphological evidence, this strongly indicates that the lobes are illuminated by cosmic ray electrons accelerated in association with star formation within this region

[1]CSIRO Astronomy and Space Science, PO Box 276, Parkes, New South Wales 2870, Australia. [2]Max-Planck-Institut für Kernphysik, PO Box 103980, 69029 Heidelberg, Germany. [3]Research School of Astronomy and Astrophysics, Australian National University, Weston Creek, Australian Capital Territory 2611, Australia. [4]International Centre for Radio Astronomy Research, M468, University of Western Australia, Crawley, Western Australia 6009, Australia. [5]ARC Centre of Excellence for All-sky Astrophysics (CAASTRO), M468, University of Western Australia, 35 Stirling Highway, Crawley, Western Australia 6009, Australia. [6]Department of Astrophysics/IMAPP, Radboud University Nijmegen, PO Box 9010, 6500 GL Nijmegen, The Netherlands. [7]Leiden Observatory, Leiden University, PO Box 9513, 2300 RA Leiden, The Netherlands. [8]Sydney Institute for Astronomy, School of Physics, The University of Sydney, New South Wales 2006, Australia. [9]Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, Massachusetts 02138, USA. [10]CSIRO Astronomy and Space Science, PO Box 76, Epping, New South Wales 1710, Australia. [11]INAF Osservatorio Astronomico di Cagliari, Strada 54 Località Poggio dei Pini, I-09012 Capoterra (CA), Italy.

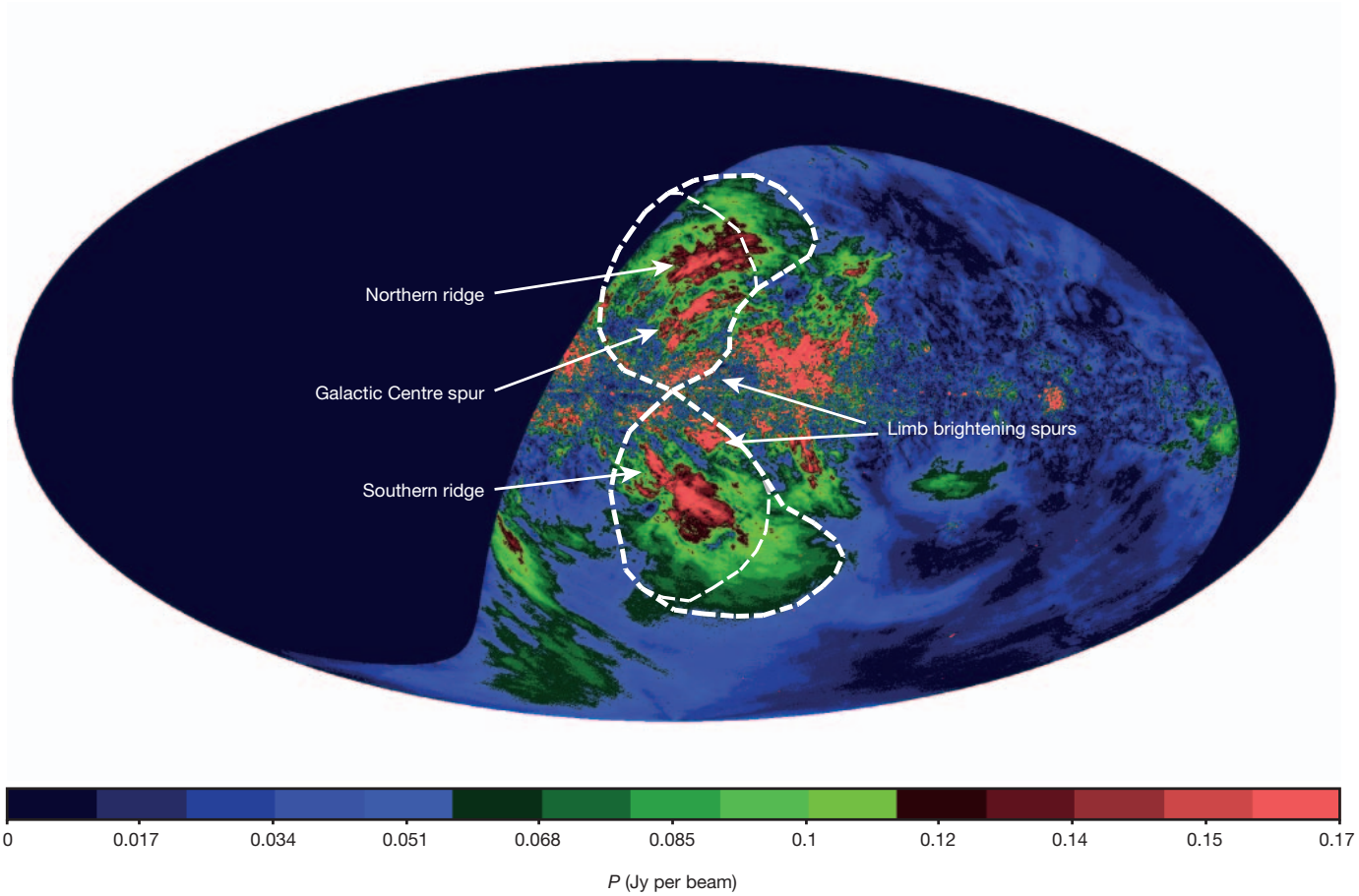


Figure 1 | Linearly polarized intensity *P* at 2.3 GHz from S-PASS. The thick dashed lines delineate the radio lobes reported in this Letter, while the thin dashed lines delimit the γ-ray Fermi bubbles[2]. The map is in Galactic coordinates, centred at the Galactic Centre with Galactic east to the left and Galactic north up; the Galactic plane runs horizontally across the centre of the map. The linearly polarized intensity flux density P (a function of the Stokes parameters Q and U,$P \equiv \sqrt{Q^2+U^2}$) is indicated by the colour scale, and given in units of Jy per beam with a beam size of 10.75′ (1 Jy $\equiv 10^{-26}$ W m^{-2} Hz^{-1}). The lobes' edges follow the γ-ray border up to Galactic latitude $b \approx |30|^\circ$, from which the radio emission extends. The three polarized radio ridges discussed in the text are also indicated, along with the two limb brightening spurs. The ridges appear to be the front side of a continuous winding of collimated structures around the general biconical outflow of the lobes (see text). The Galactic Centre spur is nearly vertical at low latitude, possibly explained by a projection effect if it is mostly at the front of the northern lobe. At its higher latitudes, the Galactic Centre spur becomes roughly parallel with the northern ridge (above), which itself exhibits little curvature; this is consistent with the overall outflows becoming cylindrical above 4–5 kpc as previously suggested[11]. In such a geometry, synchrotron emission from the rear side of each cone is attenuated by a factor $\gtrsim 2$ with respect to the front side, rendering it difficult to detect the former against the foreground of the latter and of the Galactic plane (see Supplementary Information).

(see Supplementary Information), and that the lobes are not a result of black hole activity.

The ridges appear to be continuous windings of individual, collimated structures around a general biconical outflow out of the Galactic Centre. The sense of Galactic rotation (clockwise as seen from Galactic north) and angular-momentum conservation mean that the ridges get 'wound up'[20] in the outflow with increasing distance from the plane, explaining the projected curvature of the visible, front-side of the ridges towards Galactic west. Polarized, rear-side emission is attenuated, rendering it difficult to detect against the stronger emission from the lobes' front-side and the Galactic plane (Fig. 1 and Supplementary Information).

For cosmic ray electrons synchrotron-emitting at 2.3 GHz to be able to ascend to the top of the northern ridge at ∼7 kpc in the time it takes them to cool (mostly via synchrotron emission itself) requires vertical transport speeds of >500 km s^{-1} (for a field of 15 μG; see Fig. 4). Given the geometry of the Galactic Centre spur, the outflowing plasma is moving at 1,000–1,100 km s^{-1} (Fig. 4 and Supplementary Information), somewhat faster than the ∼900 km s^{-1} gravitational escape velocity from the Galactic Centre region[21], implying that 2.3-GHz-radiating electrons can, indeed, be advected to the top of the ridges before they lose all their energy.

Given the calculated fields and the speed of the outflow, the total magnetic energy for each of the ridges, $(4–9) \times 10^{52}$ erg (see Supplementary Information), is injected at a rate of $\sim 10^{39}$ erg s^{-1} over a few million years; this is very close to the rate at which independent modelling[6] suggests Galactic Centre star formation is injecting magnetic energy into the region's outflow. On the basis of the ridges' individual energetics, geometry, outflow velocity, timescales and plasma content (see Supplementary Information), we suggest that their footpoints are energized by and rotate with the super-stellar clusters inhabiting[1] the inner ∼100 pc (in radius) of the Galaxy. In fact, we suggest that the ridges constitute 'phonographic' recordings of the past ∼10 Myr of Galactic Centre star formation. Given its morphology, the Galactic Centre spur probably still has an active footprint. In contrast, the northern and southern ridges seem not to connect to the plane at 2.3 GHz. This may indicate their footpoints are no longer active, though the southern ridge may be connected to the plane by a γ-ray counterpart (see Fig. 2). Unfortunately, present data do not allow us to trace the Galactic Centre spur all the way down to the plane: but a connection is plausible between this structure and one (or some combination) of the ∼1°-scale radio continuum spurs[15,22] emanating north of the star-forming giant molecular cloud complexes Sagittarius B and C; a connection is also plausible with the bright,

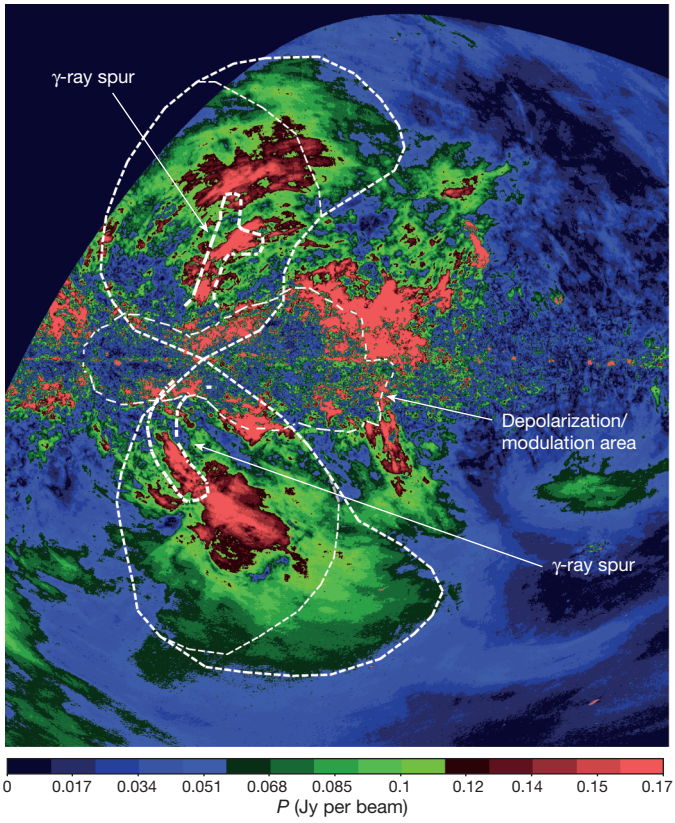


Figure 2 | Lobes' polarized intensity and γ-ray spurs. Schematic rendering of the edges of two γ-ray substructures evident in the 2–5 GeV Fermi data as displayed in figure 2 of ref. 2, which seem to be counterparts of the Galactic Centre spur and the southern ridge. The map is in Galactic coordinates, with Galactic east to the left and Galactic north up; the Galactic plane runs horizontally across the centre of the map approximately. The linearly polarized intensity flux density P is indicated by the colour scale, and given in units of Jy per beam with a beam size of 10.75′. The latter appears to be connected to the Galactic Centre by its γ-ray counterpart. With the flux densities and polarization fraction quoted in the text, we can infer equipartition[26] magnetic field intensities of $B_{eq} \approx 6\,\mu G$ ($1\,\mu G \equiv 10^{-10}$ T) if the synchrotron-emitting electrons occupy the entire volume of the lobes, or ~12 μG if they occupy only a 300-pc-thick skin (the width of the ridges). For the southern ridge, $B_{eq} \approx 13\,\mu G$; for the Galactic Centre spur, $B_{eq} \approx 15\,\mu G$; and, for the northern ridge, $B_{eq} \approx 14\,\mu G$. Note the large area of depolarization and small-angular-scale signal modulation visible across the Galactic plane extending up to $|b| \approx 10^{\circ}$ on either side of the Galactic Centre (thin dashed line). This depolarization is due to Faraday rotation by a number of shells that match Hα emission regions[27], most of them lying in the Sagittarius arm at distances from the Sun up to 2.5 kpc, and some in the Scutum-Centaurus arm at ~3.5 kpc. The small-scale modulation is associated with weaker Hα emission encompassing the same H II regions and most probably associated with the same spiral arms. Thus 2.5 kpc constitutes a lower limit to the lobes' near-side distance and places the far side beyond 5.5 kpc from the Sun (compare ref. 9). Along with their direction in the sky, this suggests that the lobes are associated with the Galactic bulge and/or Centre.

non-thermal 'radio arc'[1] (itself longitudinally coincident with the ~4-Myr-old Quintuplet[23] stellar cluster).

The magnetic energy content of both lobes is much larger than the ridges, $(1–3) \times 10^{55}$ erg. This suggests the magnetic fields of the lobes are the result of the accumulation of a number of star formation episodes. Alternatively, if the lobes' field structure were formed over the same timescale as the ridges, it would have to be associated with recent activity of the supermassive black hole, perhaps occurring in concert with enhanced nuclear star-formation activity[4].

Our data indicate that the process of gas accretion onto the Galactic nucleus inescapably involves star formation which, in turn, energizes an outflow. This carries away low-angular-momentum gas, cosmic rays and magnetic field lines, and has a number of important consequences. First, the dynamo activity in the Galactic Centre[24], probably required to generate its strong[17] *in situ* field, requires the continual expulsion of small-scale helical fields to prevent dynamo saturation[25]; the presence of the ridges high in the halo may attest to this process. Second, the lobes and ridges reveal how the very active star formation in the Galactic Centre generates and sustains a strong, large-scale magnetic field structure in the Galactic halo. The effect of this on

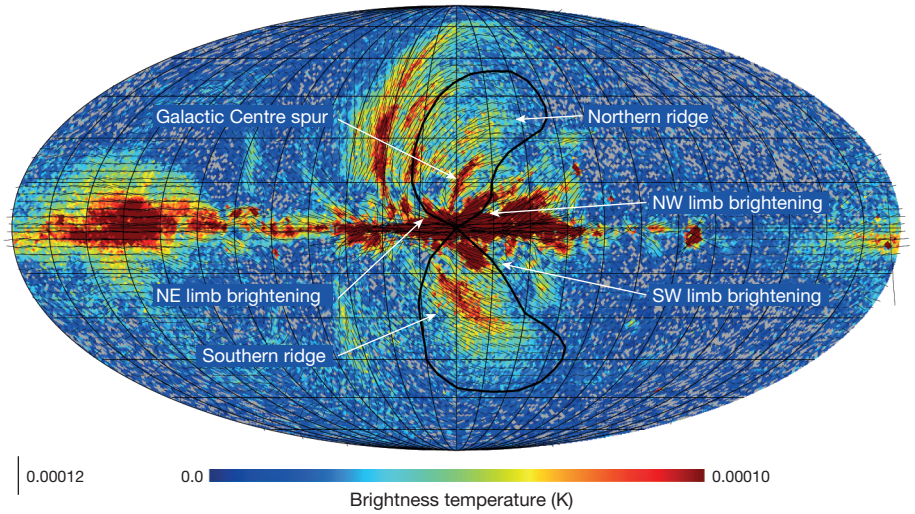


Figure 3 | Polarized intensity and magnetic angles at 23 GHz from WMAP[8]. The magnetic angle is orthogonal to the emission polarization angle and traces the magnetic field direction projected on to the plane of the sky (headless vector lines). The three ridges are obvious while traces of the radio lobes are visible (2.3 GHz edges shown by the black solid line). The magnetic field is aligned with the ridges and curves following their shape. Two spurs match the lobe edges northwest and southwest of Galactic Centre and could be limb brightening of the lobes. A third limb brightening spur candidate is also visible northeast of the Galactic Centre. The map is in Galactic coordinates, centred at the Galactic Centre. Grid lines are spaced by 15°. The emission intensity is plotted as brightness temperature, in K. The vector line length is proportional to the polarized brightness temperature (the scale is shown by the line in the bottom-left corner, in K). Data have been binned in $1^{\circ} \times 1^{\circ}$ pixels to improve the signal-to-noise ratio. From a combined analysis of microwave and γ-ray data (see also Supplementary Information) we can derive the following magnetic field limits (complementary to the equipartition limits reported in the text and Fig. 2): for the overall lobes/bubbles, $B > 9\,\mu G$; and for the Galactic Centre spur, $11\,\mu G < B < 18\,\mu G$.

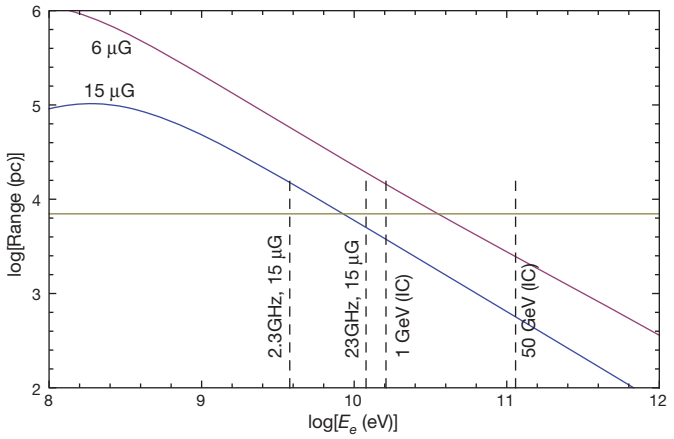


Figure 4 | The vertical range of cosmic ray electrons as a function of their kinetic energy, E_e. Two cases are reported, for field amplitudes of 15 and 6 μG (blue and red curves, respectively). Owing to geometrical uncertainties, adiabatic losses cannot be determined so the plotted range (*y* axis) actually constitutes an upper limit. Electrons are taken to be transported with a speed given by the sum of the inferred vertical wind speed (1,100 km s^{-1}) and the vertical component of the Alfven velocity in the magnetic field. The former is inferred from the geometry of the northern ridge: if its footpoint has executed roughly half an orbit in the time the Galactic Centre spur has ascended to its total height of ~4 kpc, its upward velocity must be close to[9] 1,000 km s^{-1} × $(r/100\,\mathrm{pc})^{-1}$ × $v_{\mathrm{rot}}/(80\,\mathrm{km\,s^{-1}})$, where we have normalized to a footpoint rotation speed of 80 km s^{-1} at a radius of 100 pc from the Galactic Centre[12] (detailed analysis gives 1,100 km s^{-1}: see Supplementary Information). In a strong, regular magnetic field, the electrons are expected to stream ahead of the gas at the Alfven velocity[28] in either the ridges ($B \approx 15$ μG, $v_{\mathrm{A}}^{\mathrm{vert}} \approx 300$ km s^{-1}; this is a lower limit given that $n_{\mathrm{H}} \lesssim 0.008$ cm^{-3} on the basis of the ROSAT data[29]) or in the large-scale field of the lobes ($B \approx 6$ μG, $v_{\mathrm{A}}^{\mathrm{vert}} \gtrsim 100$ km s^{-1} for $n_{\mathrm{H}} \lesssim 0.004$ cm^{-3} in the lobes' interior as again implied by the data). Also plotted as the vertical dashed lines are the characteristic energies of electrons synchrotron radiating at 2.3 and 23 GHz (for a 15 μG field) and into 1-GeV and 50-GeV γ-rays via inverse Compton ('IC') upscattering of a photon background with characteristic photon energy 1 eV; and the approximate 7 kpc distance of the top of the northern ridge from the Galactic plane.

the propagation of high-energy cosmic rays in the Galactic halo should be considered. Third, the process of gas expulsion in the outflow may explain how the Milky Way's supermassive black hole is kept relatively quiescent[1], despite sustained, inward movement of gas.

Received 8 August; accepted 26 October 2012.

1. Morris, M. & Serabyn, E. The Galactic Centre environment. *Annu. Rev. Astron. Astrophys.* **34,** 645–701 (1996).
2. Su, M., Slatyer, T. R. & Finkbeiner, D. P. Giant gamma-ray bubbles from Fermi-LAT: active galactic nucleus activity or bipolar galactic wind? *Astrophys. J.* **724,** 1044–1082 (2010).
3. Su, M. & Finkbeiner, D. P. Evidence for gamma-ray jets in the Milky Way. *Astrophys. J.* **753,** 61 (2012).
4. Zubovas, K., King, A. R. & Nayakshin, S. The Milky Way's Fermi bubbles: echoes of the last quasar outburst? *Mon. Not. R. Astron. Soc.* **415,** L21–L25 (2011).
5. Crocker, R. M. & Aharonian, F. Fermi bubbles: giant, multibillion-year-old reservoirs of Galactic Center cosmic rays. *Phys. Rev. Lett.* **106,** 101102 (2011).
6. Crocker, R. M. Non-thermal insights on mass and energy flows through the Galactic Centre and into the Fermi bubbles. *Mon. Not. R. Astron. Soc.* **423,** 3512–3539 (2012).
7. Carretti, E. in *The Dynamic ISM: A Celebration of the Canadian Galactic Plane Survey* (eds Kothes, R., Landecker, T. L. & Willis, A. G.) 276–287 (ASP Conf. Ser. CS-438, Astronomical Society of the Pacific, 2011).
8. Hinshaw, G. *et al.* Five-year Wilkinson Microwave Anisotropy Probe observations: data processing, sky maps, and basic results. *Astrophys. J.* **180** (suppl.), 225–245 (2009).
9. Jones, D. I., Crocker, R. M., Reich, W., Ott, J. & Aharonian, F. A. Magnetic substructure in the northern Fermi bubble revealed by polarized microwave emission. *Astrophys. J.* **747,** L12–L15 (2012).
10. Sofue, Y., Reich, W. & Reich, P. The Galactic center spur — A jet from the nucleus? *Astrophys. J.* **341,** L47–L49 (1989).
11. Bland-Hawthorn, J. & Cohen, M. The large-scale bipolar wind in the Galactic Center. *Astrophys. J.* **582,** 246–256 (2003).
12. Molinari, S. *et al.* A 100 pc elliptical and twisted ring of cold and dense molecular clouds revealed by Herschel around the Galactic Center. *Astrophys. J.* **735,** L33–L39 (2011).
13. Aharonian, F. A. *et al.* Discovery of very-high-energy γ-rays from the Galactic Centre ridge. *Nature* **439,** 695–698 (2006).
14. Crocker, R. M. *et al.* Wild at heart: the particle astrophysics of the Galactic Centre. *Mon. Not. R. Astron. Soc.* **413,** 763–788 (2011).
15. Law, C. J. A multiwavelength view of a mass outflow from the Galactic Center. *Astrophys. J.* **708,** 474–484 (2010).
16. Crocker, R. M. *et al.* γ-rays and the far-infrared-radio continuum correlation reveal a powerful Galactic Centre wind. *Mon. Not. R. Astron. Soc.* **411,** L11–L15 (2011).
17. Crocker, R. M., Jones, D. I., Melia, F., Ott, J. & Protheroe, R. J. A lower limit of 50 microgauss for the magnetic field near the Galactic Centre. *Nature* **463,** 65–67 (2010).
18. Yun, M. S., Reddy, N. A. & Condon, J. J. Radio properties of infrared-selected galaxies in the IRAS 2 Jy sample. *Astrophys. J.* **554,** 803–822 (2001).
19. Launhardt, R., Zylka, R. & Mezger, P. G. The nuclear bulge of the Galaxy III. Large scale physical characteristics of stars and interstellar matter. *Astron. Astrophys.* **384,** 112–139 (2002).
20. Heesen, V., Beck, R., Krause, M. & Dettmar, R.-J. Cosmic rays and the magnetic field in the nearby starburst galaxy NGC 253 III. Helical magnetic fields in the nuclear outflow. *Astron. Astrophys.* **535,** A79 (2011).
21. Muno, M. P. *et al.* Diffuse X-ray emission in a deep Chandra image of the Galactic Center. *Astrophys. J.* **613,** 326–342 (2004).
22. Pohl, M., Reich, W. & Schlickeiser, R. Synchrotron modelling of the 400 pc spur at the galactic center. *Astron. Astrophys.* **262,** 441–454 (1992).
23. Hußmann, B., Stolte, A., Brandner, W. & Gennaro, M. The present-day mass function of the Quintuplet cluster. *Astron. Astrophys.* **540,** A57 (2012).
24. Ferrière, K. Interstellar magnetic fields in the Galactic center region. *Astron. Astrophys.* **505,** 1183–1198 (2009).
25. Brandenburg, A. & Subramanian, K. Astrophysical magnetic fields and nonlinear dynamo theory. *Phys. Rep.* **417,** 1–209 (2005).
26. Beck, R. & Krause, M. Revised equipartition and minimum energy formula for magnetic field strength estimates from radio synchrotron observations. *Astron. Nachr.* **326,** 414–427 (2005).
27. Gaustad, J. E., McCullough, P. R., Rosing, W. & Van Buren, D. A robotic wide-angle Hα survey of the southern sky. *Publ. Astron. Soc. Pacif.* **113,** 1326–1348 (2001).
28. Kulsrud, R. & Pearce, W. P. The effect of wave-particle interactions on the propagation of cosmic rays. *Astrophys. J.* **156,** 445–469 (1969).
29. Almy, R. C. *et al.* Distance limits on the bright X-ray emission toward the Galactic Center: evidence for a very hot interstellar medium in the galactic X-ray bulge. *Astrophys. J.* **545,** 290–300 (2000).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work has been carried out in the framework of the S-band Polarization All Sky Survey collaboration (S-PASS). We thank the Parkes Telescope staff for support, both while setting up the non-standard observing mode and during the observing runs. R.M.C. thanks F. Aharonian, R. Beck, G. Bicknell, D. Jones, C. Law, M. Morris, C. Pfrommer, W. Reich, A. Stolte, T. Porter and H. Völk for discussions, and the Max-Planck-Institut für Kernphysik for supporting his research. R.M.C. also acknowledges the support of a Future Fellowship from the Australian Research Council through grant FT110100108. B.M.G. and C.P. acknowledge the support of an Australian Laureate Fellowship from the Australian Research Council through grant FL100100114. M.H. acknowledges the support of research programme 639.042.915, which is partly financed by the Netherlands Organisation for Scientific Research (NWO). The Parkes Radio Telescope is part of the Australia Telescope National Facility, which is funded by the Commonwealth of Australia for operation as a National Facility managed by CSIRO. We acknowledge the use of WMAP data and the HEALPix software package.

Author Contributions E.C. performed the S-PASS observations, was the leader of the project, developed and performed the data reduction package, and did the main analysis and interpretation. R.M.C. provided theoretical analysis and interpretation. L.S.-S., M.H. and S.P. performed the S-PASS observations. M.J.K. performed the telescope special set-up that allowed the survey execution. L.S.-S., M.H., B.M.G., G.B., M.J.K. and S.P. were co-proposers and contributed to the definition of the project. C.P. performed the estimate of the Hα depolarizing region distance. E.C. and R.M.C. wrote the paper together. All the authors discussed the results and commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to E.C. (Ettore.Carretti@csiro.au).

LETTER

doi:10.1038/nature11567

Optical-field-induced current in dielectrics

Agustin Schiffrin[1]†, Tim Paasch-Colberg[1], Nicholas Karpowicz[1], Vadym Apalkov[2], Daniel Gerster[3], Sascha Mühlbrandt[1,3], Michael Korbman[1], Joachim Reichert[3], Martin Schultze[1,4], Simon Holzner[1], Johannes V. Barth[3], Reinhard Kienberger[1,3], Ralph Ernstorfer[1,3,5], Vladislav S. Yakovlev[1,4], Mark I. Stockman[2] & Ferenc Krausz[1,4]

The time it takes to switch on and off electric current determines the rate at which signals can be processed and sampled in modern information technology[1–4]. Field-effect transistors[1–3,5,6] are able to control currents at frequencies of the order of or higher than 100 gigahertz, but electric interconnects may hamper progress towards reaching the terahertz (10^{12} hertz) range. All-optical injection of currents through interfering photoexcitation pathways[7–10] or photoconductive switching of terahertz transients[11–16] has made it possible to control electric current on a subpicosecond timescale in semiconductors. Insulators have been deemed unsuitable for both methods, because of the need for either ultraviolet light or strong fields, which induce slow damage or ultrafast breakdown[17–20], respectively. Here we report the feasibility of electric signal manipulation in a dielectric. A few-cycle optical waveform reversibly increases—free from breakdown—the a.c. conductivity of amorphous silicon dioxide (fused silica) by more than 18 orders of magnitude within 1 femtosecond, allowing electric currents to be driven, directed and switched by the instantaneous light field. Our work opens the way to extending electronic signal processing and high-speed metrology into the petahertz (10^{15} hertz) domain.

Three basic types of solid—metals, semiconductors and dielectrics—fundamentally differ in their reaction to an applied electric field. Metals respond to a small field with a current linearly proportional to it. This implies a strong screening that prevents high fields and charge density gradients from forming inside metals, which makes it fundamentally difficult to control their responses. By contrast, semiconductors, in which there is a relatively small but non-zero energy gap, Δ_g, between the valence and conduction bands, allow electric fields to penetrate and charge gradients to build up. This forms the basis of contemporary digital electronics[1–3,21]. Dielectrics, with their large bandgap and conduction and valence band widths, offer the fastest response. However, because of their reaction to applied fields (that is, extremely low conductivity at low fields and breakdown at high fields[17–20]), they have been thought to be unsuitable for electronic signal processing in field-effect devices. Here we demonstrate that a strong, few-cycle optical field is capable of transforming the dielectric into a state of highly increased polarizability, allowing optical currents to flow and resulting in macroscopic charge separation that is detectable in an external circuit.

In our experiments, we exposed a fused silica[22] sample ($\Delta_g \approx 9$ eV and conduction band width $\Delta_c \approx 10$ eV) to a strong, waveform-controlled, few-cycle field $F_i(t)$ with a photon energy of $\hbar\omega_L \approx 1.7$ eV (where $\hbar$ denotes Planck's constant divided by 2π and ω_L is the carrier angular frequency of the optical field), which transforms the dielectric into an optically conducting state, that is, injects carriers. We therefore refer to this field as the injection field. By carriers, we mean electrons in highly polarizable states at optical frequencies (rather than ones capable of conducting d.c. current, requiring delocalized electrons). Two unbiased gold electrodes collect the charge separated by the optical-field-induced current (Fig. 1a, Methods Summary and Supplementary Information). First, the applied external field $F_i(t) = F_0 f(t) \cos(\omega_L t + \varphi_{CE})$, with a controlled carrier-envelope phase (CEP) φ_{CE}, a sub-4-fs envelope $f(t)$ with $f(0) = 1$, and $F_0 \approx 1.7$ V Å^{-1}, was polarized perpendicularly to the metal–dielectric interface, along coordinate x, to drive the generated carriers towards the electrodes (Fig. 1a). Fig. 2 plots Q_P, the charge transferred through the ammeter shown in Fig. 1 during exposure to a single laser pulse, as a function of the change in φ_{CE} (Fig. 2a) and as a function of F_0 (Fig. 2b). The CEP of the laser pulse is shifted by $\Delta\varphi_{CE}$ on changing the

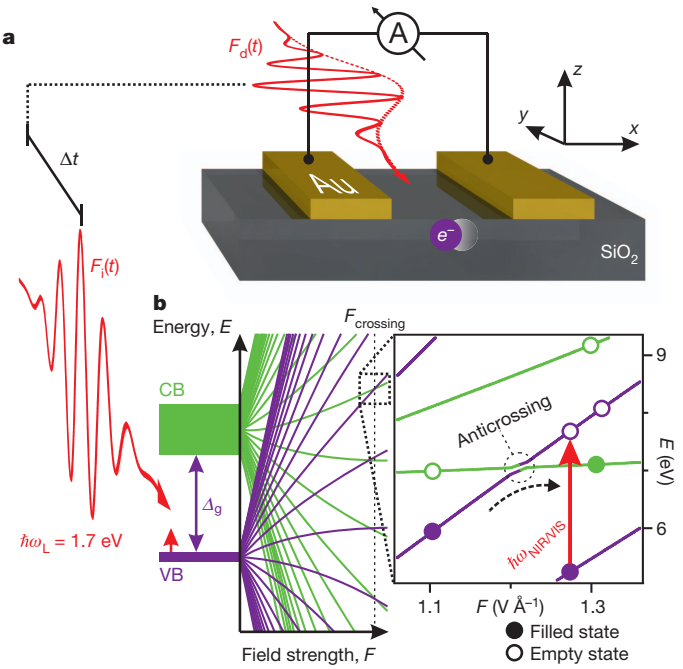


Figure 1 | Optical-field-induced conductivity and current control in a dielectric. **a**, Schematic of the metal–dielectric nanojunction. **b**, Schematic illustration of the adiabatic energy levels of the electronic states in the valence band (VB, purple) and conduction band (CB, green) of the dielectric under the influence of a static or slowly varying strong electric field. The eigenenergies fan out as the strength of the electric field increases, resulting in avoided crossings, that is, anticrossings (inset). At low fields, the valence band states are occupied, and the conduction band states are empty. As the field strength increases, the valence band and conduction band levels cross, but the respective Wannier–Stark states are localized at distant sites, and the anticrossings are passed diabatically (that is, the conduction band states remain unpopulated) until the field approaches or exceeds ~1 V Å^{-1}. At these field strengths, electrons may be promoted into the conduction band via Zener tunnelling, leaving the electron in the lower-energy state after the passage of the anticrossing (adiabatic transition, depicted by dashed arrow). The resultant unoccupied valence band states mediate strong single-photon resonances at visible/near-infrared angular frequencies $\omega_{NIR/VIS}$ within the valence band (red arrow). The emergence of these resonances results in a strong transient polarizability.

[1]Max-Planck-Institut für Quantenoptik, Hans-Kopfermann-Strasse 1, D-85748 Garching, Germany. [2]Department of Physics and Astronomy, Georgia State University, Atlanta, Georgia 30340, USA. [3]Physik-Department, Technische Universität München, James-Franck-Strasse, D-85748 Garching, Germany. [4]Fakultät für Physik, Ludwig-Maximilians-Universität, Am Coulombwall 1, D-85748 Garching, Germany. [5]Fritz-Haber-Institut der Max-Planck-Gesellschaft, Faradayweg 4–6, 14195 Berlin, Germany. †Present addresses: Department of Physics and Astronomy, University of British Columbia, Vancouver, British Columbia, V6T 1Z1 Canada; Quantum Matter Institute, University of British Columbia, Vancouver, British Columbia, V6T 1Z4 Canada.

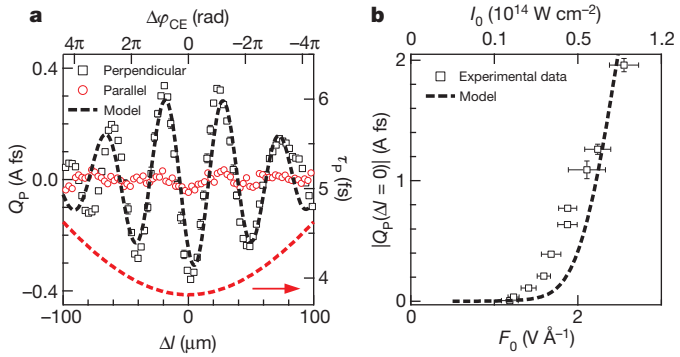


Figure 2 | Carrier-envelope-phase control and intensity dependence of optical-field-generated electric current in SiO_2. **a**, Plot of the φ_{CE}-dependent component of $Q_P(\Delta l)$ against the change, Δl, in the propagation length through a fused silica wedge with respect to the propagation length yielding the minimum pulse duration ($\Delta l = 0$), for polarizations perpendicular (along x; squares) and parallel (in the y–z plane; circles) to the metal–dielectric interface ($F_0 \approx 1.7$ V Å^{-1}). The data represent an average of several consecutively acquired values at a given value of Δl. Error bars show the standard deviation, and most are smaller than the size of the symbols representing the mean values of the data. The black dashed line shows the prediction of our quantum mechanical model. For the parallel polarization, the signal is suppressed by more than an order of magnitude. The residual signal is attributed to microscopic imperfections of the macroscopically plane metal–dielectric interface, which result in locations with a non-zero perpendicular component of the field. The red dashed line (right axis) depicts the change in pulse duration τ_P (full width at half maximum intensity) as a function of Δl, taking into account the group velocity dispersion of the visible/near-infrared pulse in the fused silica wedge. **b**, Plot of the maximum amplitude, $|Q_P(\Delta l = 0)|$, of the transferred charge against the peak amplitude, F_0, of the applied external field polarized along x: measurement (squares) and theoretical prediction (dashed line). For a given F_0 value, $|Q_P(\Delta l = 0)|$ is determined by fitting the most pronounced oscillation of $Q_P(\Delta l)$ (**a**, squares) with a sine function. The vertical error bars account for the standard deviation of such fit. Different values of F_0 correspond to different beam sizes. Data points have been normalized accordingly. F_0 is determined by monitoring the pulse energy and the laser beam waist at the focus, and the horizontal error bars quantify random fluctuations in these parameters. I_0 is the corresponding peak intensity of the optical pulse. Data from our model have been multiplied by the effective cross-section, A_{eff}, of the metal–dielectric interfaces confining the active volume of the dielectric (main text).

propagation length by Δl in a pair of thin, fused silica wedges[23] (Methods Summary and Supplementary Fig. 1).

We find that Q_P varies quasi-periodically with Δl. Its minimum at $\Delta l = 0$ turns into a maximum at $\Delta l = \pm \Delta l_{deph}$, with $\Delta l_{deph} = 23$ μm shifting[23] the CEP by $\Delta\varphi_{CE} = \pi$, which results in an approximate inversion of the optical waveform: $F_i(t, \Delta l_{deph}) \approx -F_i(t, 0)$. This reversal of Q_P and its order-of-magnitude reduction when $F_i(t)$ is polarized parallel to the metal–dielectric interface (that is, in the y–z plane) provide the first significant indication that this current is induced directly by the instantaneous light field. The maximum value of Q_P scales nearly exponentially with the field amplitude and reaches ~10^4 electrons before breakdown (Fig. 2b). This nonlinear dependence is responsible for the decay of $Q_P(\Delta l)$ with increasing Δl, as a consequence of dispersive pulse broadening.

To decouple the injection and driving processes, we irradiated the hybrid junction with two synchronized, orthogonally polarized fields. An injection field with a peak amplitude of $F_0^{(i)} \approx 2$ V Å^{-1} was polarized parallel to the electrode–insulator interface (that is, in the y–z plane) to prevent it from driving current through the circuit. This driving was accomplished by a weaker version of the same field, the drive field $F_d(t)$, which had an amplitude of $F_0^{(d)} \approx 0.2$ V Å^{-1} (Fig. 1a). The delay, Δt, between the peaks of $F_i(t)$ and $F_d(t)$ determines the timing of carrier injection with respect to the drive field and thereby controls the momentum that $F_d(t)$ transfers to the charge carriers. Fig. 3a shows Q_P versus Δt for the CEPs of the injection ($\varphi_{CE}^{(i)}$) and drive ($\varphi_{CE}^{(d)}$) pulses corresponding to one of the zero-crossings of Q_P near $\Delta l = 0$ in Fig. 2a. Such values of $\varphi_{CE}^{(i)}$ and $\varphi_{CE}^{(d)}$ are chosen to avoid a residual (background) current originating in any of the fields independently (Fig. 2a).

The transferred charge periodically changes its sign as a function of Δt, revealing two major current reversal cycles equal to the field oscillation cycle of $F_d(t)$ (Fig. 3a). The magnitude of the signal rapidly decays outside this few-femtosecond delay interval. Fig. 3b plots the result of the same measurement performed with an 'inverted' drive field, in which the field oscillations are reversed by shifting the CEP by $\Delta\varphi_{CE}^{(d)} = \pi$. For any value of Δt, the transferred charge is reversed with respect to its value at the same delay in Fig. 3a. This is exactly the behaviour we intuitively expect for a current driven directly by $F_d(t)$. The observed $Q_P(\Delta t)$ reflecting the oscillating behaviour of $F_d(t)$ provides compelling evidence that the light field governs the current that emerges from the dielectric medium and is measured in the external circuit. The transferred charge was also found to depend sensitively on the CEP of the injection field $\varphi_{CE}^{(i)}$ (Supplementary Fig. 7). This also corroborates the evidence that carrier injection is induced directly by the instantaneous field $F_i(t)$ rather than by effects governed by the cycle-averaged intensity. Experiments on free-standing electrodes kept under helium flux did not yield any φ_{CE}-dependent electronic signal under identical irradiation conditions (Supplementary Fig. 10), and so we identify the insulator as the primary source of current in our experiments.

Fig. 3c plots the electric field waveform of the drive pulse retrieved from attosecond streaking[24]. Such streaking experiments were performed in a parallel ultrahigh vacuum set-up independently from the light-field-induced current measurements and under identical laser conditions. The good correspondence between the solid lines in Fig. 3a and Fig. 3c suggests that the injected carriers sample the optical field with good fidelity. Carrier injection must therefore be substantially confined to a half-period, that is, to a time window of ~1 fs. This confinement is also indicative of a strong nonlinearity, in accordance with Fig. 2b.

To describe our observations theoretically, we calculate the current density, $j_x(t)$, generated in the bulk insulator along coordinate x by solving the time-dependent Schrödinger equation for a dielectric film with a thickness in the range 50–500 nm exposed to a strong optical field. We neglect the Coulomb interaction between the electrons, which is justified by the extremely short (~1 fs) timescale on which the processes occur. The time dependence of the applied electric fields is taken from attosecond-streaking experiments (Fig. 3c). We compute $j_x(t)$ as the temporal derivative of the field-induced polarization, multiplying it by the effective cross-section, A_{eff}, of the metal–dielectric interfaces confining the active volume of the dielectric. More details are given in the Supplementary Information.

Our computations predict the dependence of

$$Q_P = A_{eff} \int_{-\infty}^{\infty} j_x(t)\, dt$$

on the CEP (Fig. 2a, black dashed line), the peak electric field (Fig. 2b), and the delay between the injection and drive fields for different $F_d(t)$ (Fig. 3) and $F_i(t)$ (Supplementary Fig. 7) waveforms. All the observed dependencies have been excellently reproduced using parameters—defining the band structure and three relevant dipole matrix elements—that we chose once and did not subsequently adjust. The parameter A_{eff} was adjusted to fit the magnitude of the collected charge, and $A_{eff} \approx 5 \times 10^{-12}$ m^2 gave the best agreement with the measurements. This value matches the experimentally estimated metal–dielectric cross-section within an order of magnitude. Hence, our microscopic model predicts our observations both qualitatively and, within an order of magnitude, quantitatively. This quantitative agreement also implies that a significant fraction of the charge separated by the optical currents in the dielectric is transferred to the metal leads, in

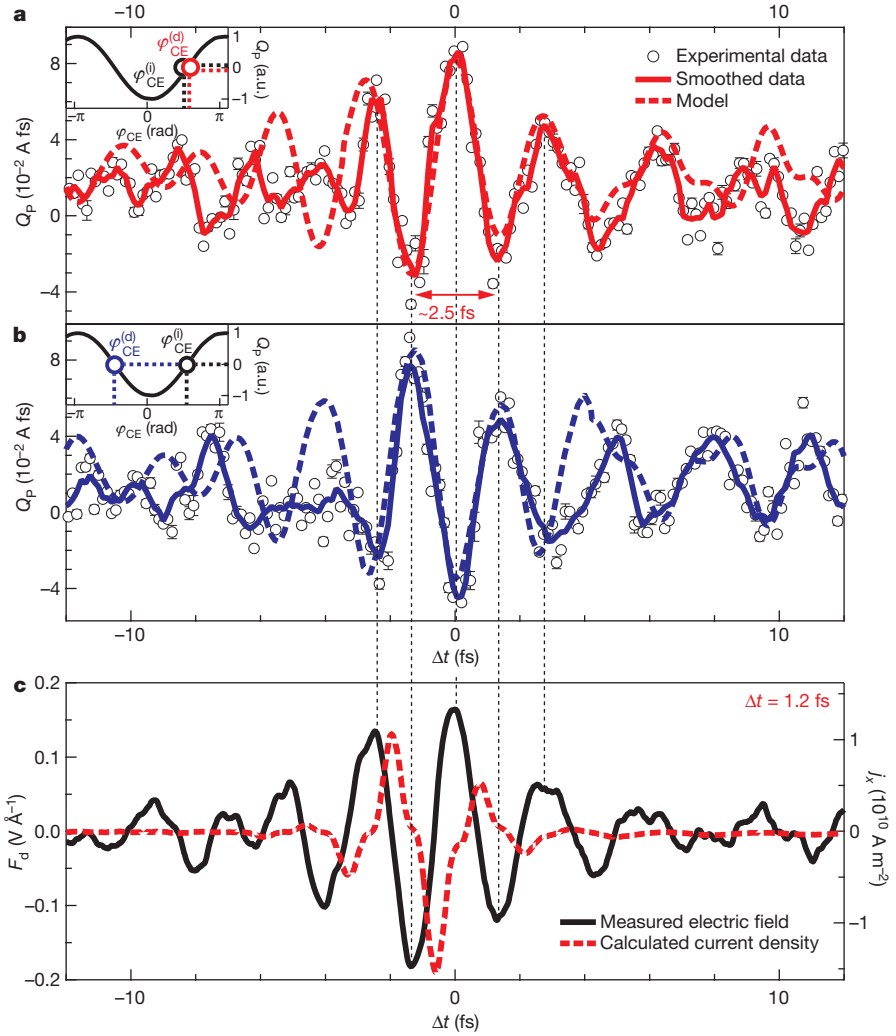


Figure 3 | Subfemtosecond control of electric current with the electric field of light. **a**, **b**, Transferred charge, Q_P, versus delay, Δt, between the injection ($F_i(t)$) and drive ($F_d(t)$) fields, with parameters as described in the text. For $\Delta t < 0$, $F_d(t)$ precedes $F_i(t)$. The CE phases of the injection and drive fields, $\varphi_{CE}^{(i)}$ and $\varphi_{CE}^{(d)}$ respectively, are set such that Q_P induced independently by any of the two fields vanishes (insets). The experimental data shown were recorded under identical conditions except for $\varphi_{CE}^{(d)}$, which is shifted by π between the two measurements (compare the insets). The experimental data (circles) represent an average of several consecutively acquired values at a given Δt value. The error bars are the respective standard deviations. The solid lines correspond to the smoothed data obtained by averaging over adjacent points. The dashed lines show the prediction of our microscopic quantum mechanical model. Data from our model have been multiplied by the effective cross-section, A_{eff}, of the metal–dielectric interfaces confining the active volume of the dielectric (main text). The slightly asymmetric polarity of the experimental signal could be due to the non-zero residual current for each individual pulse. The dependence of the current on Δt demonstrates that the charge-carrier generation takes place predominantly in the silica medium and that electrons of the gold electrodes, which are not sensitive to the component of the field parallel to the metal–insulator interface, that is, $F_i(t)$, do not contribute to the optically injected current. **c**, Real-time optical electric field of the visible/near-infrared pulses retrieved from attosecond-streaking measurements[24] (solid black line), which were performed under identical laser conditions. The magnitude of the field is normalized to the strength of the drive pulse, $F_d(t)$, used in the photocurrent experiment. The red dashed curve shows the time-dependent current density, $j_x(t)$, as calculated from our quantum mechanical model for $\Delta t = 1.2$ fs. The vertical black dashed lines provide a guide to the eye for comparing the temporal evolution of the measured electronic signal with that of the drive electric field.

line with the predictions of a theoretical study of optically induced currents in a single-molecule nanojunction[10] (Supplementary Information, section 3.4).

The calculated $j_x(t)$ in Fig. 3c and Supplementary Fig. 3 reveals ultrafast turn-on and turn-off behaviour, allowed by the extremely wideband energy spectrum of the dielectric. This predicted temporal behaviour enables us to evaluate the peak current density from our measurements. With a maximum value of $Q_P(\Delta t)$ on the order of 0.1 A fs (Fig. 3a, b) and A_{eff} as specified above, our experimental data yield a transferred charge density of the order of 10^{10} A fs m^{-2}. Given that the duration of the current density pulse is of the order of 1 fs (see $j_x(t)$ in Figs 3c and Supplementary Fig. 3), we estimate a peak current density perpendicular to the electrodes of $J_x \approx 10^{10}$ A m^{-2}, driven by a peak field amplitude of $F_0^{(d)} \approx 0.2 \times 10^{10}$ V m^{-1}. We may now introduce an effective electrical conductivity at optical frequencies as $\sigma_{eff}(\omega_L) = J_x / F_0^{(d)} \approx 5\ \Omega^{-1}$ m^{-1}, which exceeds the static d.c. conductivity of amorphous silica, $\sigma_{a-SiO_2} < 10^{-18}\ \Omega^{-1}$ m^{-1}, by more than 18 orders of magnitude.

Our microscopic theory correctly predicts the field-induced increase in polarizability that is responsible for the observed current, but does not provide intuitive insight into its origin. We suggest, with reference to a detailed discussion in the Supplementary Information, that this field-enhanced polarizability can be qualitatively interpreted in terms of the dynamic formation of Wannier–Stark states[25–28] localized at a certain site, l, of the lattice along the direction of the applied field, F, with an energy of $E_l = E_n - eaFl$ (e is the electron charge, a is

the lattice constant and E_n is the band offset) in the adiabatic limit (Fig. 1b). For high fields, $F > 1\,\mathrm{V\,\AA^{-1}}$, the eigenenergies of the Wannier–Stark states from the valence band and conduction band that reside at neighbouring lattice sites $|l_v\rangle$ and $|l_c\rangle$ become equal; see the avoided crossing and anticrossing of the respective purple and green plots of energy against field the inset of Fig. 1b. For increasing field strength, adiabatic passage of such an anticrossing via Zener-type tunnelling[17] may occur with a significant probability (Fig. 1b inset, dashed arrow), emptying a valence band state. The resultant unoccupied valence band states give rise to strong visible/near-infrared single-photon resonances within the valence band (Fig. 1b inset, red arrow). The emergence of these resonances directly implies a strong transient polarizability[29,30].

This field-induced transformation is predicted to be reversible and highly nonlinear (as suggested by the resultant $j_x(t)$ in Fig. 3c and Supplementary Fig. 3). The consequence is a sub-cycle increase in the polarizability of the system, which leads to an asymmetric charge displacement along the field vector: the averaging of the current to zero that is inherent in linear processes is eliminated. Consequently, net charge accumulations of opposite sign—dependent on the waveform (Fig. 2) or on the delay between injection and driving fields (Fig. 3)—form at the opposite facets of the dielectric and transfer to the electrodes.

Our experiments reveal an ultrafast 'turning on' of the measured current. At the same time, they do not provide direct evidence for the similarly fast 'turning off' predicted by our theory. A proof of the ultrafast turn-off behaviour of the underlying field-induced nonlinear polarization and conduction band population was provided by recent time-resolved absorption and reflection experiments[31]. This suggests that conductivity and, consequently, current can be switched on and off in a dielectric using optical fields on a timescale of less than, or of the order of, 1 fs, and that this can be done without incurring much dissipation. This operation cycle is similar to that occurring in a field-effect transistor, but is fundamentally more energy efficient: when switching, the field-effect transistor dissipates by electron–hole recombination all energy stored in it, whereas the dielectric in our experiment returns almost all stored energy to the injection field. These possibilities may have ramifications for overcoming the speed limits of semiconductor electronics. In the shorter term, our work holds promise for the development of a solid-state device for direct sampling of electric transients with bandwidths extending to optical frequencies.

METHODS SUMMARY

Source of strong ultrashort optical waveforms. A customized titanium–sapphire chirped-pulse amplifier produces linearly polarized, visible/near-infrared laser pulses with controllable φ_{CE}, a pulse energy of ~400 μJ and a repetition rate of 3 kHz. The ultrabroad spectrum of the laser spans the 450–1,100-nm spectral range, supporting a sub-4-fs pulse duration (full width at half maximum of the intensity temporal profile), which corresponds to a ~1.5-cycle pulse at the carrier wavelength of ~750 nm. The laser pulses can be focused onto the investigated solid-state system such that the cycle-averaged peak intensity reaches $10^{14}\,\mathrm{W\,cm^{-2}}$ (see Supplementary Fig. 2 for properties of the laser).

Photoactive metal–dielectric nanocircuit. The photoactive circuit on which the experiments have been performed consists of a metal–dielectric–metal nanojunction. This structure is fabricated by cleaving amorphous SiO_2 (Crystec GmbH) and coating the adjacent surfaces of the atomically sharp cleaved silica edge with ~50 nm of gold evaporated at grazing incidence. We obtain regular, straight, metal electrodes isolated from each other by dielectric nanoscale trenches with widths of the order of ~50 nm. No voltage bias is applied to the electrodes during the experiment. The junction is coated with a supplementary sputtered silica nanofilm such that the metal electrodes are embedded in a homogeneous silica matrix. See Supplementary Information for more details on the sample structure and experimental set-up. Identical measurements were made on samples composed of cleaved monocrystalline SiO_2(001) (Crystec GmhH) and on flat, ~500-nm gold–SiO_2–gold nanogaps. Both yielded results very similar to those reported in this work, but it was possible to distinguish possible effects arising from the lack of long-range order in the amorphous insulator and from local plasmonic resonances at the metal–dielectric interface. These points are discussed in Supplementary Information.

Detection scheme for the optically induced φ_{CE}-dependent electric signal. In this work, we present experimental data on the φ_{CE}-dependent component of the optically generated electronic signals in the circuit. We isolate the electric current induced by the electric field of the optical waveform from the φ_{CE}-independent contributions (for example electric currents caused by photo-ionization of gold due to imbalanced irradiation of the electrodes). The observable of interest is filtered by synchronizing the pulse train from the chirped-pulse amplifier in such a way that two consecutive ultrashort light bursts have a φ_{CE}-change of π. This results in a modulation of φ_{CE} at half the repetition rate of the pulse train. The optically generated electronic signal in the nanocircuit is measured using a current–voltage converter with a variable high gain and a bandwidth supporting the laser pulse train. The φ_{CE}-dependent component of the current is extracted with a lock-in amplifier locked at the φ_{CE} modulation frequency. The presented data correspond to experiments performed in ambient atmosphere at ~20 °C. We carried out identical measurements in vacuum, yielding the same results.

Received 20 December 2011; accepted 31 August 2012.
Published online 5 December 2012.

1. Kahng, D. Electric field controlled semiconductor device. US patent 3,102 230 (1963).
2. Taur, Y. & Ning, T. H. *Fundamentals of modern VLSI devices* (Cambridge Univ. Press, 1998).
3. Liou, J. J. & Schwierz, F. *Modern Microwave Transistors: Theory, Design and Performance* (Wiley-Interscience, 2003).
4. Caulfield, H. J. & Dolev, S. Why future supercomputing requires optics. *Nature Photon.* **4,** 261–263 (2010).
5. Schwierz, F. & Liou, J. J. RF transistors: recent developments and roadmap toward terahertz applications. *Solid-State Electron.* **51,** 1079–1091 (2007).
6. Deal, W. R. *et al.* Low noise amplification at 0.67 THz using 30 nm InP HEMTs. *IEEE Microw. Wirel. Compon. Lett.* **21,** 368–370 (2011).
7. Kurizki, G., Shapiro, M. & Brumer, P. Phase-coherent control of photocurrent directionality in semiconductors. *Phys. Rev. B* **39,** 3435–3437 (1989).
8. Atanasov, R., Hache, A., Hughes, J. L. P., van Driel, H. M. & Sipe, J. E. Coherent control of photocurrent generation in bulk semiconductors. *Phys. Rev. Lett.* **76,** 1703–1706 (1996).
9. Prechtel, L. *et al.* Time-resolved picosecond photocurrents in contacted carbon nanotubes. *Nano Lett.* **11,** 269–272 (2011).
10. Franco, I., Shapiro, M. & Brumer, P. Robust ultrafast currents in molecular wires through stark shifts. *Phys. Rev. Lett.* **99,** 126802 (2007).
11. Valley, G. C. Photonic analog-to-digital converters. *Opt. Express* **15,** 1955–1982 (2007).
12. Nagatsuma, T. Photonic measurement technologies for high-speed electronics. *Meas. Sci. Technol.* **13,** 1655–1663 (2002).
13. Auston, D. H. Picosecond optoelectronic switching and gating in silicon. *Appl. Phys. Lett.* **26,** 101–103 (1975).
14. Auston, D. H. Ultrafast optoelectronics. *Top. Appl. Phys.* **60,** 183–233 (1988).
15. Shimosato, H., Ashida, M., Itoh, T., Saito, S. & Sakai, K. in *Ultrafast Optics V* (eds Watanabe, S. & Midorikawa, K.) 317–323 (Springer Ser. Opt. 132, Springer, 2007).
16. Katzenellenbogen, N. & Grischkowsky, D. Efficient generation of 380 fs pulses of THz radiation by ultrafast laser-pulse excitation of a biased metal-semiconductor interface. *Appl. Phys. Lett.* **58,** 222–224 (1991).
17. Zener, C. A theory of the electrical breakdown of solid dielectrics. *Proc. R. Soc. Lond. A* **145,** 523–529 (1934).
18. Rethfeld, B. Free-electron generation in laser-irradiated dielectrics. *Phys. Rev. B* **73,** 035101 (2006).
19. Jones, S. C., Braunlich, P., Casper, R. T., Shen, X. A. & Kelly, P. Recent progress on laser-induced modifications and intrinsic bulk damage of wide-gap optical-materials. *Opt. Eng.* **28,** 1039–1068 (1989).
20. Lenzner, M. *et al.* Femtosecond optical breakdown in dielectrics. *Phys. Rev. Lett.* **80,** 4076–4079 (1998).
21. Schwierz, F., Wong, H. & Liou, J. J. *Nanometer CMOS* (Pan Stanford, 2010).
22. Koslowski, T., Kob, W. & Vollmayr, K. Numerical study of the electronic structure of amorphous silica. *Phys. Rev. B* **56,** 9469–9476 (1997).
23. Xu, L. *et al.* Route to phase control of ultrashort light pulses. *Opt. Lett.* **21,** 2008–2010 (1996).
24. Kienberger, R. *et al.* Atomic transient recorder. *Nature* **427,** 817–821 (2004).
25. Wannier, G. H. Wave functions and effective Hamiltonian for Bloch electrons in an electric field. *Phys. Rev.* **117,** 432–439 (1960).
26. Bleuse, J., Bastard, G. & Voisin, P. Electric-field-induced localization and oscillatory electro-optical properties of semiconductor superlattices. *Phys. Rev. Lett.* **60,** 220–223 (1988).
27. Mendez, E. E., Agulló-Rueda, F. & Hong, J. M. Stark localization in GaAs-GaAlAs superlattices under an electric field. *Phys. Rev. Lett.* **60,** 2426–2429 (1988).
28. Bar-Joseph, I. *et al.* Room-temperature electroabsorption and switching in a GaAs/AlGaAs superlattice. *Appl. Phys. Lett.* **55,** 340–342 (1989).
29. Durach, M., Rusina, A., Kling, M. F. & Stockman, M. I. Metallization of nanofilms in strong adiabatic electric fields. *Phys. Rev. Lett.* **105,** 086803 (2010).
30. Durach, M., Rusina, A., Kling, M. F. & Stockman, M. I. Predicted ultrafast dynamic metallization of dielectric nanofilms by strong single-cycle optical fields. *Phys. Rev. Lett.* **107,** 086602 (2011).
31. Schultze, M. *et al.* Controlling dielectrics with the electric field of light. *Nature* doi:10.1038/nature11720 (this issue).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank P. Altpeter and Y. Deng for technical support and discussions, and we thank the Munich-Centre for Advanced Photonics for financial support. A.S. acknowledges the Alexander von Humboldt Foundation and the Swiss National Science Foundation. N.K. acknowledges the Alexander von Humboldt Foundation. The work of M.I.S. and V.A. was supported by the Chemical Sciences, Biosciences and Geosciences Division (grant no. DEFG02-01ER15213) and by the Materials Sciences and Engineering Division (grant no. DE-FG02-11ER46789) of the Office of the Basic Energy Sciences, Office of Science, US Department of Energy. R.K. acknowledges an ERC Starting Grant.

Author Contributions A.S., R.K., R.E. and F.K. designed and supervised the experiments. A.S., T.P.-C., D.G., S.M., J.R. and J.V.B. participated in sample design and fabrication. A.S., T.P.-C., N.K., D.G., S.M., M.S. and S.H. performed the measurements. A.S., N.K., V.A., M.K., V.S.Y. and M.I.S. took part in the theoretical modelling. A.S., T.P.-C., N.K., R.K., R.E., V.S.Y. and F.K. analysed and interpreted the experimental data. All authors discussed the results and contributed to the final manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to A.S. (aschiffr@phas.ubc.ca), M.I.S. (mstockman@gsu.edu) or F.K. (ferenc.krausz@mpq.mpg.de).

LETTER

doi:10.1038/nature11720

Controlling dielectrics with the electric field of light

Martin Schultze[1,2], Elisabeth M. Bothschafter[1,3], Annkatrin Sommer[1], Simon Holzner[1], Wolfgang Schweinberger[1], Markus Fiess[1], Michael Hofstetter[2], Reinhard Kienberger[1,3], Vadym Apalkov[4], Vladislav S. Yakovlev[2], Mark I. Stockman[4] & Ferenc Krausz[1,2]

The control of the electric and optical properties of semiconductors with microwave fields forms the basis of modern electronics, information processing and optical communications. The extension of such control to optical frequencies calls for wideband materials such as dielectrics, which require strong electric fields to alter their physical properties[1–5]. Few-cycle laser pulses permit damage-free exposure of dielectrics to electric fields of several volts per ångström[6] and significant modifications in their electronic system[6–13]. Fields of such strength and temporal confinement can turn a dielectric from an insulating state to a conducting state within the optical period[14]. However, to extend electric signal control and processing to light frequencies depends on the feasibility of reversing these effects approximately as fast as they can be induced. Here we study the underlying electron processes with sub-femtosecond solid-state spectroscopy, which reveals the feasibility of manipulating the electronic structure and electric polarizability of a dielectric reversibly with the electric field of light. We irradiate a dielectric (fused silica) with a waveform-controlled near-infrared few-cycle light field of several volts per angström and probe changes in extreme-ultraviolet absorptivity and near-infrared reflectivity on a timescale of approximately a hundred attoseconds to a few femtoseconds. The field-induced changes follow, in a highly nonlinear fashion, the turn-on and turn-off behaviour of the driving field, in agreement with the predictions of a quantum mechanical model. The ultrafast reversibility of the effects implies that the physical properties of a dielectric can be controlled with the electric field of light, offering the potential for petahertz-bandwidth signal manipulation.

A dielectric subjected to a weak optical field reacts to its change instantly (adiabatically) as long as the laser frequency $\omega_L \ll \Delta_{gap}/\hbar$, where Δ_{gap} is the gap between the valence band and conduction band; for silica $\Delta_{gap} \approx 9$ eV (numerical values below are given for this material). When the strength of the electric field F approaches the critical field strength, inducing a change in electron potential energy by Δ_{gap} over the lattice period $a \approx 5$ Å then

$$F_{crit} = \frac{\Delta_{gap}}{|e|a} \approx 2\,\mathrm{V\,\AA^{-1}} \quad (1)$$

(where e is the electron charge), and Zener-type transitions[1] inject electrons into the entire conduction band within its spectral width $\Delta_c \approx 10$ eV. This by itself, irrespective of ω_L, leads to ultrafast dynamics of the induced broadband polarization on a timescale of $\hbar/\Delta_c \approx 0.1$ fs. Hence, real-time access to strong-field-induced dynamics in dielectrics calls for sub-femtosecond temporal resolution. In our work, this was provided by the envelope of sub-100-attosecond extreme-ultraviolet (XUV) pulses and the controlled instantaneous field of few-femtosecond near-infrared (NIR) laser pulses.

In a first set of experiments, the waveform-controlled linearly polarized field, $F_L(t)$, of NIR laser pulses of less than 4 fs (carrier wavelength $\lambda_L = 780$ nm and wave cycle $T_L = 2\pi/\omega_L = 2.6$ fs), along with isolated 72-as XUV pulses carried at $\hbar\omega_{XUV} \approx 105$ eV and polarized parallel to the laser field, impinged collinearly on a free-standing 125-nm SiO_2 film (see Fig. 1a and Methods). The XUV pulse probed the processes induced by $F_L(t)$ by promoting electrons from the L-shell of silicon into the conduction band states. Transient changes in the conduction band are reflected in the XUV spectra recorded as a function of delay of the XUV probe with respect to the NIR-field excitation; see Fig. 1c. We note that the XUV pulse induces polarization over the entire band $\Delta_c/\hbar$, which largely dephases within $\tau_{dephasing} \sim \hbar/\Delta_c \sim 100$ as after passage of the XUV pulse through the sample. The XUV light radiated from the sample coherently with the XUV probe field therefore decayed within about 100 as after the XUV pulse. This permitted—in contrast with previous experiments in atomic gases[15]—us to record attosecond transient absorption spectra that provide information about the evolving state of the quantum system at the instant of probing and are not appreciably affected by the ongoing strong-field excitation afterwards.

We supplemented attosecond probing with simultaneous streaking[16] of the excitation field $F_L(t)$ (Fig. 1b) in a neon gas jet in front of the SiO_2 sample. Figure 1a displays the laser field and the XUV pulse envelope retrieved from this measurement. Simultaneous implementation of these two attosecond techniques permitted assignment of the transient absorption spectra to well-defined moments within the excitation field, $F_L(t)$, providing absolute timing of the processes under scrutiny.

For a detailed analysis, we evaluate the absorbance A, given by $A(\hbar\omega_{XUV}) = \alpha(\hbar\omega_{XUV})d$, where $\alpha(\hbar\omega_{XUV})$ is the absorption coefficient at the photon energy $\hbar\omega_{XUV}$, and d is the thickness of the sample. We first discuss XUV absorption near the edge of the conduction band. The solid line in Fig. 2b shows the transient change in A induced by the NIR field plotted in Fig. 2a; its measured peak electric field strength is $F_0 = 2.5 \pm 0.5$ V Å^{-1}. It reveals a large (>10%) increase of the sample's XUV transmittance over sub-femtosecond intervals. The oscillations in absorbance at a frequency of $2\omega_L$ are accompanied by similar oscillations in the position of the conduction-band edge (Fig. 2c). The measurement resolves condensed-matter processes driven by the instantaneous field of visible light. The changes emerge within a single half-cycle of the driving field and terminate with similar abruptness several femtoseconds later; see also the inset of Fig. 2b. The XUV absorption bleaching extends, with similar temporal behaviour, over the entire conduction band; see Supplementary Fig. 1.

To analyse the observed field-induced modifications, we adapted—with exactly the same parameters—the quantum-mechanical model developed for describing electric currents in dielectrics induced by strong, few-cycle NIR fields under physical conditions identical to the experiments of this work[14]. Briefly, it is based on the time-dependent Schrödinger equation in the presence of the measured field $F_L(t)$, incorporating field screening but neglecting electron collisions and excitonic effects. For a detailed description, see Supplementary Information. The model accurately predicts the magnitude and the sub-femtosecond temporal structure of the observed modulation in strength (Fig. 2b) and position (Fig. 2c) of the 109-eV absorption line, as well as the bandwidth and the strength of the XUV absorption bleaching (Supplementary Fig. 1). Agreement in the phase of these modulations with respect to the oscillating laser field (Fig. 2b, c)

[1]Max-Planck-Institut für Quantenoptik, Hans-Kopfermann-Strasse 1, D-85748 Garching, Germany. [2]Fakultät für Physik, Ludwig-Maximilians-Universität, Geschwister-Scholl-Platz 1, D-80539 München, Germany. [3]Physik-Department, Technische Universität München, James-Franck-Strasse, D-85748 Garching, Germany. [4]Department of Physics, Georgia State University, Atlanta, Georgia 30340, USA.

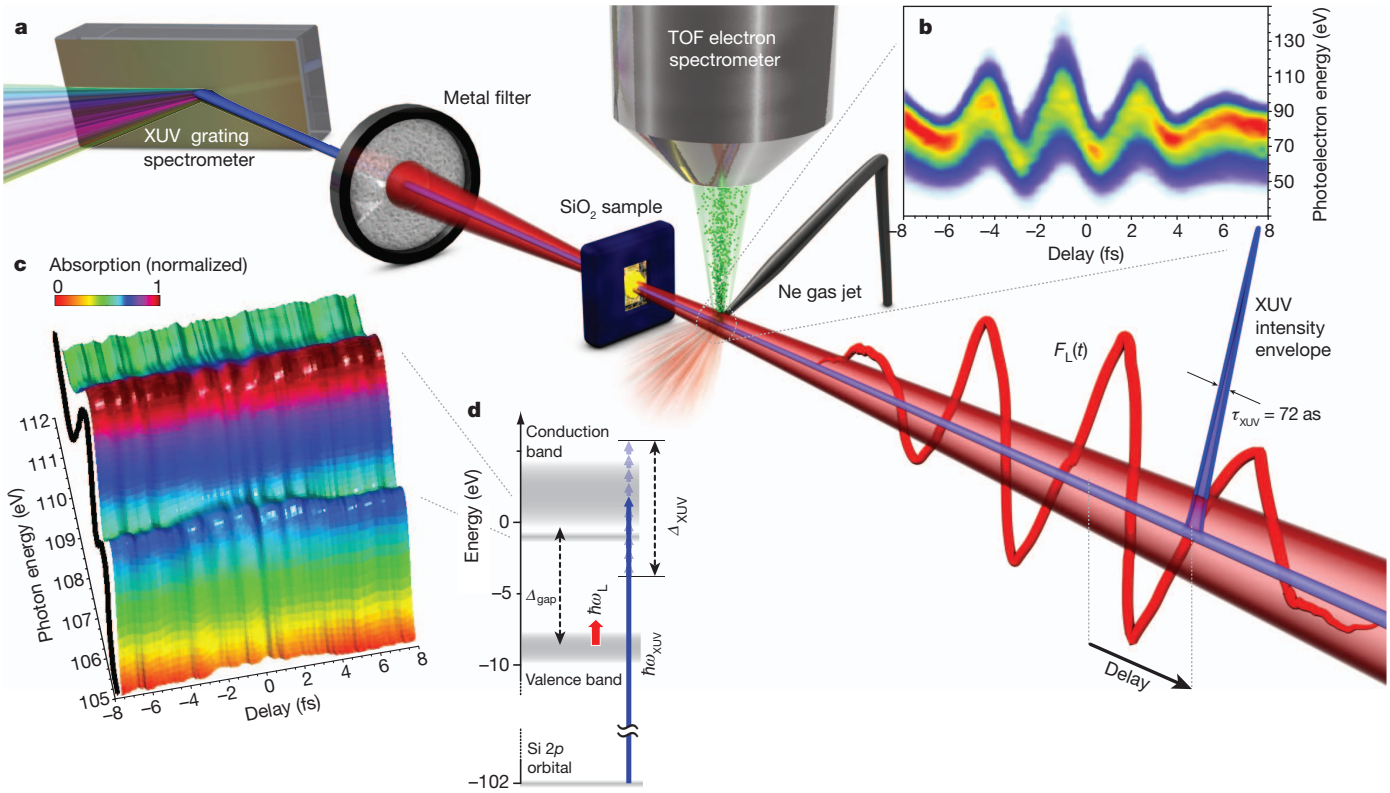


Figure 1 | Simultaneous attosecond absorption and streaking spectroscopy. **a**, Schematic illustration of the experimental set-up. TOF, time of flight. **b**, Attosecond streaking spectrogram of the 72-attosecond (full-width at half maximum) isolated XUV pulse and a near-single-cycle NIR pulse used in the experiments. The spectrogram was recorded by measuring the energy spectrum of photoelectrons released by the XUV pulse into the NIR field as a function of NIR–XUV delay. **c**, Attosecond transient absorption spectrogram of a 125-nm-thick SiO_2 membrane in the range 105–112 eV as a function of delay between the sub-100-attosecond XUV pulse and the NIR laser pulse with a delay step of $\Delta t_{delay} = 100$ as (the black line shows the synchrotron-XANES absorption data[24] for comparison). **d**, Schematic energy level diagram of SiO_2. The blue arrow represents the XUV absorption from Si L-states to the SiO_2 conduction band. The 9-eV bandgap between valence and conduction band exceeds the NIR photon energy by a factor of 6, as indicated by the red arrow.

supports our understanding that it is the instantaneous field that drives the processes.

In a second set of experiments, we scrutinized the influence of the strong field on the polarizability in the NIR range by studying the transmitted and reflected NIR light (for details, see Methods and Supplementary Information). Within our experimental accuracy of about $\pm 1.2\%$ (dictated by the laser pulse-to-pulse stability), the transmittance of our NIR beam did not vary up to the intensity threshold for breakdown. In contrast, the power of the reflected beam, detected in a *z*-scan geometry, increased by nearly an order of magnitude for field strengths approaching the breakdown threshold as shown in Fig. 3a. To ensure the highest contrast for the subsequent pump–probe study (see below), this measurement was implemented with

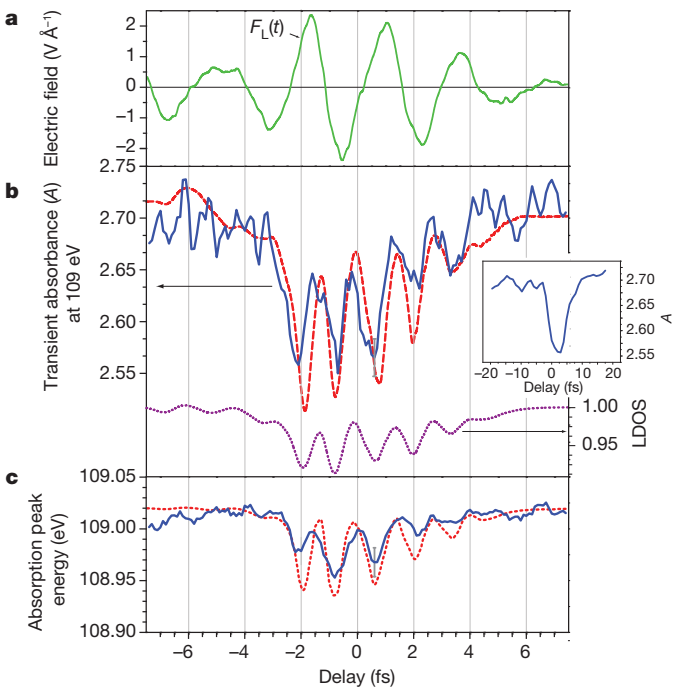


Figure 2 | Attosecond time-resolved strong-field-induced effects in SiO_2. Solid lines are experimental results; dashed lines are predictions of theoretical modelling. **a**, Electric field of the few-cycle NIR laser pulse impinging on the SiO_2 sample, $F_L(t)$, as extracted from attosecond streaking (see Fig. 1b). **b**, Transient change of the absorbance $A(\hbar\omega_{XUV}) = \alpha(\hbar\omega_{XUV})d$ integrated over a 1-eV bandwidth at $\hbar\omega = 109$ eV, as a function of the delay (the delay step is $\Delta t_{delay} = 100$ as) between the 72-as XUV probe and the NIR laser pulse (blue solid line), along with the prediction of our quantum mechanical model (red dashed line). The inset shows the evolution of A in a more extended delay range, recorded with larger delay steps ($\Delta t_{delay} = 0.5$ fs). The error bar in **b** represents the standard error of the average over 15 spectral lineouts within the energy range 108.5–109.5 eV; it exhibits little variation over the delay scan. The dashed violet line is the calculated local density of states (LDOS) at the position of a Si atom (integrated over the energy range accessed by the XUV pulse, for more details, see Supplementary Information) versus delay of the XUV probe. **c**, Energy of the absorption peak at 109 eV subject to an optical-field-induced (dynamic Stark) shift (the blue solid line shows the measurement, the red dashed line shows the calculation). The error bar in **c** represents the 95% confidence interval of the peak position of a least-squares fit of the area of a Gaussian function to the area under the absorption line; it exhibits little variation over the delay scan.

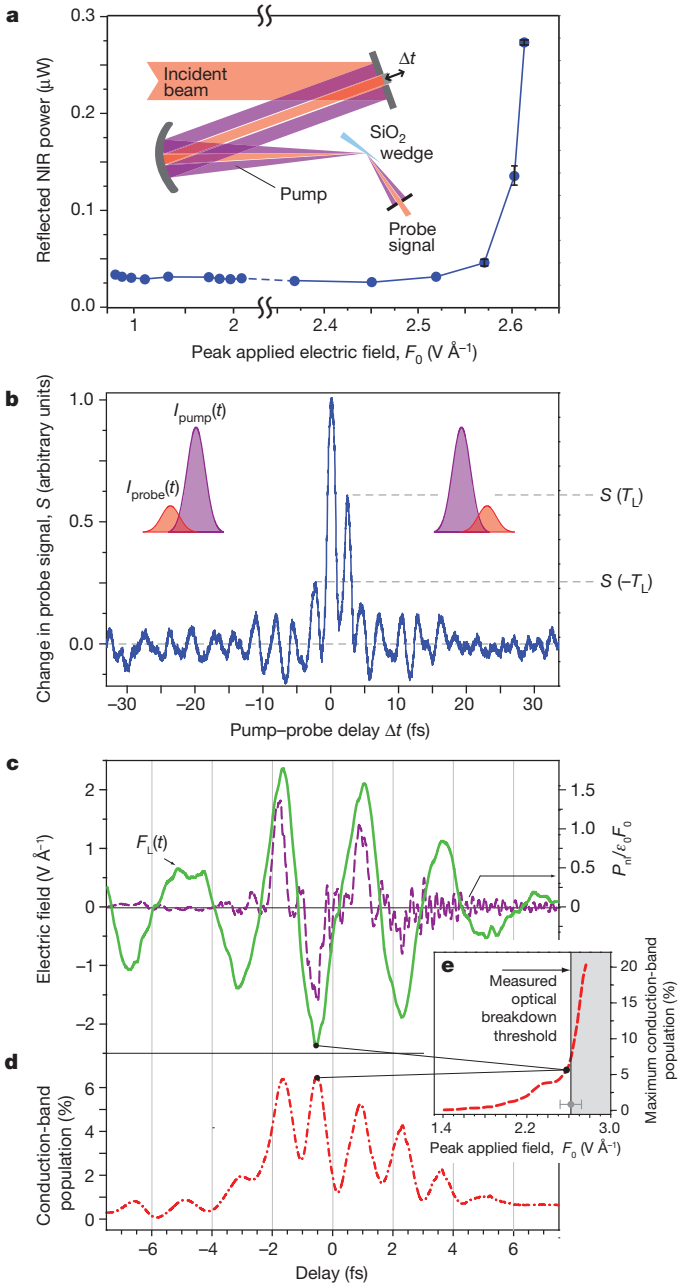


Figure 3 | Wave-cycle-resolved NIR femtosecond probing of strong-field-induced nonlinear reflectivity of SiO_2. **a**, Reflected power of *p*-polarized NIR laser pulses of less than 4 fs incident at Brewster's angle on the thin-wedged sample of fused silica as a function of peak electric field (penetrating into the sample). The process is fully reversible for several thousand laser shots before irreversible damage occurs owing to self-focusing inside the sample rather than on the surface. The data points are averaged over 3,000 laser pulses (error bars represent the standard deviation). Inset, schematic illustration of the pump–probe transient reflectivity set-up (for description, see Methods). **b**, Reflected power on axis of the probe pulse of less than 4 fs (represented by the red beam; intensity envelope I_{probe}) as a function of the delay with respect to the strong pump pulse of less than 4 fs (illustrated by the violet beam; intensity envelope I_{pump}). The curve is the result of an average of ten scans; for representative individual scans and details, see Supplementary Fig. 2 and the related description in the Supplementary Information. The reflected probe signal at delays of $\Delta t = \pm T_L$ is denoted by $S(\pm T_L)$. **c**, Electric field of the few-cycle NIR laser pulse impinging on the SiO_2 sample, $F_L(t)$, as extracted from attosecond streaking (see green solid line in Fig. 1b), and the resulting nonlinear polarization $P_{nl}(t)$, normalized to $\varepsilon_0 F_0$, predicted by our model (purple dashed line). **d**, Computed transient evolution of the population of (unperturbed) conduction band states timed to the laser field. **e**, Peak transient conduction band population versus peak applied field strength. The grey shading represents the experimental value of the optical breakdown threshold, as determined from the reflectance measurement yielding the data shown in **a**.

light impinging at Brewster's angle and polarized in the plane of incidence (*p*-polarized). The dependence on the field strength was extremely nonlinear, closely resembling that of the field-induced current[14]. Our model is not yet applicable to an oblique angle of incidence but did correctly predict the field-strength-dependence of the current[14]. The similarity in the intensity scaling of the optical current and reflectivity points to their common physical origin: the field-enhanced polarizability.

Is this effect mainly due to a conduction-band population surviving the few-cycle field or to reversible effects being not only turned on but also turned off by the field? Although the XUV bleaching already suggests that the latter is more likely, we sought additional evidence using time-resolved optical reflectometry. To this end, we split the incident pulse into a strong pump and a weak probe beam, with the sum of the two field maxima being slightly below the breakdown threshold, and measured the reflected probe intensity as a function of the delay Δt; see the sketch of the apparatus in Fig. 3a and the result in Fig. 3b.

Comparing the reflected probe signal S at delays $\Delta t = \pm T_L$, we found the signal to be significantly greater in the immediate aftermath of the strong pulse, $S(T_L)/S(-T_L) = 2.1 \pm 0.2$ (with no measurable dependence on the carrier-envelope phase to within the measurement accuracy limited by intensity fluctuations). However, comparing signals on opposite fringe peaks at $\Delta t \approx \pm 2T_L$, $\pm 3T_L$,..., showed no appreciable asymmetry—$S(nT_L) \approx S(-nT_L)$, $n > 1$—within our experimental precision. If a significant fraction of the induced nonlinear reflectivity (and, consequently, polarization) survived the pulse peak over several cycles, this would imply $S(nT_L)$ significantly larger than $S(-nT_L)$, owing to a much larger fraction of the probe pulse experiencing the increased reflectivity. Our observation thus provides compelling evidence, in a model-independent manner, for the substantial reversibility of the strong-field-induced increase in polarizability on the timescale of the optical wave cycle.

Figure 3c contrasts the applied electric field $F_L(t)$ (solid line) with the calculated field-induced nonlinear polarization $P_{nl}(t)$ normalized to $\varepsilon_0 F_0$ (dashed line). The strongly sharpened half cycles of $P_{nl}(t)$ point to an extremely nonlinear scaling with driving field strength, consistent with the data in Fig. 3a. The nonlinear response is largely confined to the central laser cycle and disappears completely before the end of the pulse. This effect is accompanied by a field-induced transient population and subsequent depopulation of the conduction band, confined, again, to several femtoseconds (see Fig. 3d). For field strengths beyond $F_0 \approx 2.5\ \mathrm{V\,Å^{-1}}$, Fig. 3e shows a near-exponential growth of conduction band population, causing breakdown. In fact, the maximum field we could safely apply to our samples before breakdown was $F_0 \approx 2.5\ \mathrm{V\,Å^{-1}}$. Thus, our model also correctly predicts the threshold for optical breakdown (Fig. 3e). Both the nonlinear polarization and the conduction band population induced by the strong field return to near-zero immediately after the laser pulse for $F_0 \leq 2.5\ \mathrm{V\,Å^{-1}}$. This is fully consistent with the abrupt decay of field-induced transient NIR reflectivity and XUV absorption bleaching: see Figs 3b and 2b, respectively. These results imply that the sample exposed to fields as high as $F_0 \leq 2.5\ \mathrm{V\,Å^{-1}}$ resumes its original (field-free) state immediately after exposure. This points towards the reversible, Hamiltonian nature of the laser-induced dynamics up to the critical field strength.

Our quantum mechanical model identifies light-field-induced reversible changes of the local density of states (Fig. 2b), the conduction-band population, and the polarizability as the mechanisms responsible for transient XUV absorption bleaching, NIR reflectivity enhancement, optical breakdown and optical-field-induced currents[14]. The model accounts for all these phenomena with a set of parameters adjusted in ref. 14, which we did not modify in this work. The ability

to predict such a large aggregate of disparate physical effects properly without parameter adjustment provides convincing evidence of the theory's validity.

Considerations presented in ref. 14 provide a hint towards a possible intuitive picture to explain the above effects in terms of the well-known phenomenon of Wannier–Stark localization[17–21]. The field-induced strong transient localization of states is expected to occur in all bands, allowing XUV-induced transitions predominantly between states localized at the same site of the lattice. The energy levels of such colocalized Wannier–Stark states of different bands are shifted identically by the field, leaving the XUV transition energies (between the L-band and conduction band) nearly unchanged. Hence, XUV probing is insensitive to these energy shifts, apart from the small dynamic shifts observed in Fig. 2c, but uncovers dynamic changes in the density of states. The Wannier–Stark ladder formation delivers an intuitive account for the computed decrease of the local density of states (Fig. 2b), which appears to be mainly responsible for the decrease of the XUV absorption. The field-enhanced polarizability of the system can also be related to the strong Stark shifts in the conduction band and valence band: they pull interband transition frequencies into resonance with the visible–NIR spectrum of the driving field[14]. The resultant enhancement in polarizability appears to be the common origin of the optical-field-induced reflectivity observed here and the currents observed in ref. 14. In spite of the consistency of this argument, the intuitive Wannier–Stark picture remains hypothetical until directly verified by experiment, for example, by probing valence band-conduction band transitions with sub-femtosecond resolution.

In conclusion, the synergistic use of attosecond streaking and absorption spectroscopy and wave-cycle-resolved optical reflectivity measurements provide real-time insight into strong-field phenomena in dielectrics on sub-femtosecond to few-femtosecond timescales. The experiments reveal, in agreement with quantum-mechanical modelling, that field-induced changes of all the physical quantities studied are reversible up to the critical field strength, that is, they can be turned on and off on the timescale of the optical period. Controlling the properties of dielectrics with the instantaneous electric field of light points the way towards petahertz signal sampling and processing technologies.

METHODS SUMMARY

XUV transmittivity. Time-resolved XUV absorption spectroscopy was performed with broadband, isolated attosecond XUV pulses generated via high harmonic generation of visible–infrared laser pulses of less than 4 fs (FemtoPower Compact Pro, Femtolasers) in neon. The collinear laser and XUV beam were separated into two arms of a Mach–Zehnder-type interferometer. The relative timing between the two arms could be adjusted with a mirror moved by a piezo translator, and a variable aperture controls the infrared intensity on target. For details, see the Supplementary Information and ref. 22. The XUV and NIR beams were focused with a toroidal mirror on free-standing chemical-vapour-deposited SiO_2 samples (either 125 nm or 250 nm thick). The XUV beam transmitted through the sample was spectrally dispersed by a flat-field grating and projected on a XUV-sensitized camera[23]. Our measurements were performed near (to within $<$10%) optical breakdown. NIR-field-induced breakdown manifested itself as macroscopic mechanical damage of the sample, which became immediately visible in the magnified image of the focal region, which was permanently monitored.

NIR transmittivity and reflectivity. Field-induced changes in NIR transmittivity through 125-nm films and the reflectivity of thin ($\sim$0.2 mm) wedged plates were studied with laser pulses of less than 4 fs gently focused and the intensity on the sample was varied by translating it longitudinally along the optical axis (z-scan). The pump–probe measurement is sketched schematically in Fig. 3a. The incident NIR beam was split into two collinear beams by a two-component plane mirror consisting of a 5-mm-diameter inner section and a concentric external section. The inner and outer sections reflected the weak probe and the strong pump beam with a reflectivity of 10% and 97.5%, respectively. The energy of the pump pulse was controlled by an aperture. The probe pulse was delayed by translating the inner section with a piezo stage and focused along with the pump pulse onto the sample. The probe beam was isolated with an iris in front of the detector measuring the reflected probe pulse energy as a function of pump–probe delay. A more detailed description of the Methods is given in the Supplementary Information.

Received 20 December 2011; accepted 24 October 2012.
Published online 5 December 2012.

1. Zener, C. A theory of the electrical breakdown of solid dielectrics. *Proc. R. Soc. Lond. A* **145,** 523–529 (1934).
2. Wannier, G. Wave functions and effective Hamiltonian for Bloch electrons in an electric field. *Phys. Rev.* **117,** 432–439 (1960).
3. Franz, W. Einfluß eines elektrischen Feldes auf eine optische Absorptionskante. *Z. Naturforsch. A* **13,** 484 (1958).
4. Keldysh, L. V. Behavior of non-metallic crystals in strong electric fields. *Sov. J. Exp. Theor. Phys.* **6,** 763 (1958).
5. Mizumoto, Y., Kayanuma, Y., Srivastava, A., Kono, J. & Chin, A. H. Dressed-band theory for semiconductors in a high-intensity infrared laser field. *Phys. Rev. B* **74,** 045216 (2006).
6. Lenzner, M. *et al.* Femtosecond optical breakdown in dielectrics. *Phys. Rev. Lett.* **80,** 4076–4079 (1998).
7. Gertsvolf, M., Spanner, M., Rayner, D. M. & Corkum, P. B. Demonstration of attosecond ionization dynamics inside transparent solids. *J. Phys. At. Mol. Opt. Phys.* **43,** 131002 (2010).
8. Mitrofanov, A. *et al.* Optical detection of attosecond ionization induced by a few-cycle laser field in a transparent dielectric material. *Phys. Rev. Lett.* **106,** 147401 (2011).
9. Shih, T., Winkler, M. T., Voss, T. & Mazur, E. Dielectric function dynamics during femtosecond laser excitation of bulk ZnO. *Appl. Phys. A* **96,** 363–367 (2009).
10. Ghimire, S. *et al.* Redshift in the optical absorption of ZnO single crystals in the presence of an intense midinfrared laser field. *Phys. Rev. Lett.* **107,** 167407 (2011).
11. Ghimire, S. *et al.* Observation of high-order harmonic generation in a bulk crystal. *Nature Phys.* **7,** 138 (2011).
12. Durach, M., Rusina, A., Kling, M. & Stockman, M. Metallization of nanofilms in strong adiabatic electric fields. *Phys. Rev. Lett.* **105,** 086803 (2010).
13. Durach, M., Rusina, A., Kling, M. & Stockman, M. Predicted ultrafast dynamic metallization of dielectric nanofilms by strong single-cycle optical fields. *Phys. Rev. Lett.* **107,** 086602 (2011).
14. Schiffrin, A. *et al.* Optical-field-induced current in dielectrics. *Nature* doi:10.1038/nature11567 (this issue).
15. Goulielmakis, E. *et al.* Real-time observation of valence electron motion. *Nature* **466,** 739–743 (2010).
16. Kienberger, R. *et al.* Atomic transient recorder. *Nature* **427,** 817–821 (2004).
17. Bloch, F. Über die Quantenmechanik der Elektronen in Kristallgittern. *Z. Phys.* **52,** 555–600 (1929).
18. Wannier, G. H. *Elements of Solid State Theory. Elements* 173–177 (Cambridge Univ. Press, 1959).
19. Bleuse, J., Bastard, G. & Voisin, P. Electric-field-induced localization and oscillatory electro-optical properties of semiconductor superlattices. *Phys. Rev. Lett.* **60,** 220–223 (1988).
20. Mendez, E., Agulló-Rueda, F. & Hong, J. Stark localization in GaAs-GaAlAs superlattices under an electric field. *Phys. Rev. Lett.* **60,** 2426–2429 (1988).
21. Mendez, E. E. & Bastard, G. Wannier-Stark ladders and Bloch oscillations in superlattices. *Phys. Today* **46,** 34, http://dx.doi.org/10.1063/1.881353 (1993).
22. Fiess, M. *et al.* Versatile apparatus for attosecond metrology and spectroscopy. *Rev. Sci. Instrum.* **81,** 093103 (2010).
23. Schultze, M. *et al.* State-of-the-art attosecond metrology. *J. Electron Spectrosc. Relat. Phenom.* **184,** 68–77 (2011).
24. Li, D. *et al.* X-ray absorption spectroscopy of silicon dioxide (SiO_2) polymorphs; the structural characterization of opal. *Am. Mineral.* **79,** 622–632 (1994).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work was supported by the Max Planck Society and the Deutsche Forschungsgemeinschaft Cluster of Excellence: Munich Centre for Advanced Photonics (www.munich-photonics.de).The work of M.I.S. and V.A. was supported by grant number DEFG02-01ER15213 from the Chemical Sciences, Biosciences and Geosciences Division and by grant number DE-FG02-11ER46789 from the Materials Sciences and Engineering Division of the Office of the Basic Energy Sciences, Office of Science, US Department of Energy. We thank K. Yabana, R. Ernstorfer and N. Karpowicz for discussions.

Author Contributions M.S., R.K., M.I.S. and F.K. conceived and supervised the study. M.S., E.M.B., A.S., S.H., W.S., M.F. and M.H. prepared and performed the experiment. V.A. and M.I.S. accomplished the theoretical modelling. M.S., E.M.B., A.S., V.S.Y. and F.K. analysed and interpreted the experimental data. All authors discussed the results and contributed to the final manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.S. (martin.schultze@mpq.mpg.de), M.I.S. (mstockman@gsu.edu) and F.K. (krausz@lmu.de).

LETTER

doi:10.1038/nature11787

Probabilistic cost estimates for climate change mitigation

Joeri Rogelj[1,2], David L. McCollum[2], Andy Reisinger[3], Malte Meinshausen[4,5] & Keywan Riahi[2,6]

For more than a decade, the target of keeping global warming below 2 °C has been a key focus of the international climate debate[1]. In response, the scientific community has published a number of scenario studies that estimate the costs of achieving such a target[2–5]. Producing these estimates remains a challenge, particularly because of relatively well known, but poorly quantified, uncertainties, and owing to limited integration of scientific knowledge across disciplines[6]. The integrated assessment community, on the one hand, has extensively assessed the influence of technological and socio-economic uncertainties on low-carbon scenarios and associated costs[2–4,7]. The climate modelling community, on the other hand, has spent years improving its understanding of the geophysical response of the Earth system to emissions of greenhouse gases[8–12]. This geophysical response remains a key uncertainty in the cost of mitigation scenarios but has been integrated with assessments of other uncertainties in only a rudimentary manner, that is, for equilibrium conditions[6,13]. Here we bridge this gap between the two research communities by generating distributions of the costs associated with limiting transient global temperature increase to below specific values, taking into account uncertainties in four factors: geophysical, technological, social and political. We find that political choices that delay mitigation have the largest effect on the cost–risk distribution, followed by geophysical uncertainties, social factors influencing future energy demand and, lastly, technological uncertainties surrounding the availability of greenhouse gas mitigation options. Our information on temperature risk and mitigation costs provides crucial information for policy-making, because it clarifies the relative importance of mitigation costs, energy demand and the timing of global action in reducing the risk of exceeding a global temperature increase of 2 °C, or other limits such as 3 °C or 1.5 °C, across a wide range of scenarios.

We generate cost distributions by combining mitigation cost estimates of emissions scenarios with probabilistic temperature projections. Importantly, our cost estimates do not account for any avoided climate damages as a result of emission reductions. This information is obtained from a large set of scenarios created with an integrated assessment model[14,15], for which the temperature increase is computed with a probabilistic climate model[16,17] (Fig. 1, Supplementary Fig. 1, Methods and Supplementary Information). Each modelling framework has inherent limitations. For example, although it incorporates state-of-the-art uncertainty quantifications of the Earth system, our model does not fully explore tipping points. Similarly our energy-economic emissions scenarios map a wide range of possible futures (Supplementary Figs 7 and 8) but are not exhaustive of all potential outcomes (Supplementary Information).

Temperature projections for any given pathway have a spread owing to geophysical uncertainties[18] (Fig. 1b). In the absence of any serious mitigation efforts (present global carbon prices of less than US$1 per tonne of carbon-dioxide-equivalent emissions (tCO_2e^{-1})), the likelihood of limiting warming to less than 2 °C is essentially zero (<1%; Fig. 1c). However, imposing a carbon price of about US$20 tCO_2e^{-1} in our model would increase the probability of staying below 2 °C to about 50%, and carbon prices of more than US$40 tCO_2e^{-1} would achieve the 2 °C objective with a probability of more than 66% ('likely' by the definition of the Intergovernmental Panel on Climate Change[19]). Similar trends hold for other cost metrics (Supplementary Information). For example, a carbon price of US$20–40 tCO_2e^{-1} translates in our model to cumulative discounted mitigation costs (2012–2100) of the order of 0.8–1.3% of gross world product (Supplementary Fig. 10).

A marked feature of the mitigation cost distribution (Fig. 2) is that the probability of global warming staying below 2 °C levels off at high carbon prices. This occurs because, beyond a given carbon price, nearly all mitigation options that can substantially influence emissions in the medium term have been deployed in our model. Higher carbon prices help further to reduce emissions later in the century, but only affect temperatures after peaking[20]. Hence, the probability of staying below 2 °C during the twenty-first century reaches an asymptote.

Geophysical uncertainties shed light on only one aspect of mitigation costs, however. To gain insight into how assumptions regarding technological and social uncertainties influence our cost distribution, we create a large set of sensitivity cases (Table 1), in which we vary some salient features of the scenarios, namely the availability and use of specific mitigation technologies; future social development and, by extension, global energy demand; and the international political context surrounding climate mitigation action, specifically delays in the implementation of a globally comprehensive mitigation response[7] (Supplementary Information). We note that population and economic growth do not vary in our scenarios; we therefore cannot assess their relative importance with our ensemble (Supplementary Information). Given its policy relevance[21], we focus most of our discussion on the limit of 2 °C (Supplementary Figs 4 and 5 illustrate the results for 2.5 and 3 °C, respectively).

Our results can be framed in two ways (Fig. 2): first, in terms of how probabilities for achieving the 2 °C objective change for a fixed cost (black arrows); and, second, in terms of how the cost consistent with the 2 °C goal varies for a given probability level (orange arrows). Whether or not a carbon price of about US$40 tCO_2e^{-1} restricts global warming to less than 2 °C with a likelihood of more than 66% depends on the future availability of key mitigation technologies (Fig. 2a). In our worst-case technology-sensitivity assumption—that capture and geological storage of carbon (CCS) is entirely unavailable—the probability of staying below 2 °C at a carbon price of US$40 tCO_2e^{-1} decreases to around 50%. However, with no such constraints and further technological breakthroughs (Table 1), the likelihood of limiting warming to 2 °C could be higher than 66% at the same carbon price.

The cost distributions also show how changes in technological measures affect the economics of mitigation given a fixed probability level. For example, in most cases the 2 °C objective can be achieved with a probability of more than 66% as long as the carbon price is high

[1]Institute for Atmospheric and Climate Science, ETH Zurich, Universitätstrasse 16, CH-8092 Zürich, Switzerland. [2]International Institute for Applied Systems Analysis (IIASA), Schlossplatz 1, A-2361 Laxenburg, Austria. [3]New Zealand Agricultural Greenhouse Gas Research Centre, Private Bag 11008, Palmerston North 4442, New Zealand. [4]School of Earth Sciences, University of Melbourne, Victoria 3010, Australia. [5]PRIMAP Group, Potsdam Institute for Climate Impact Research, PO Box 60 12 03, 14412 Potsdam, Germany. [6]Graz University of Technology, Inffeldgasse, A-8010 Graz, Austria.

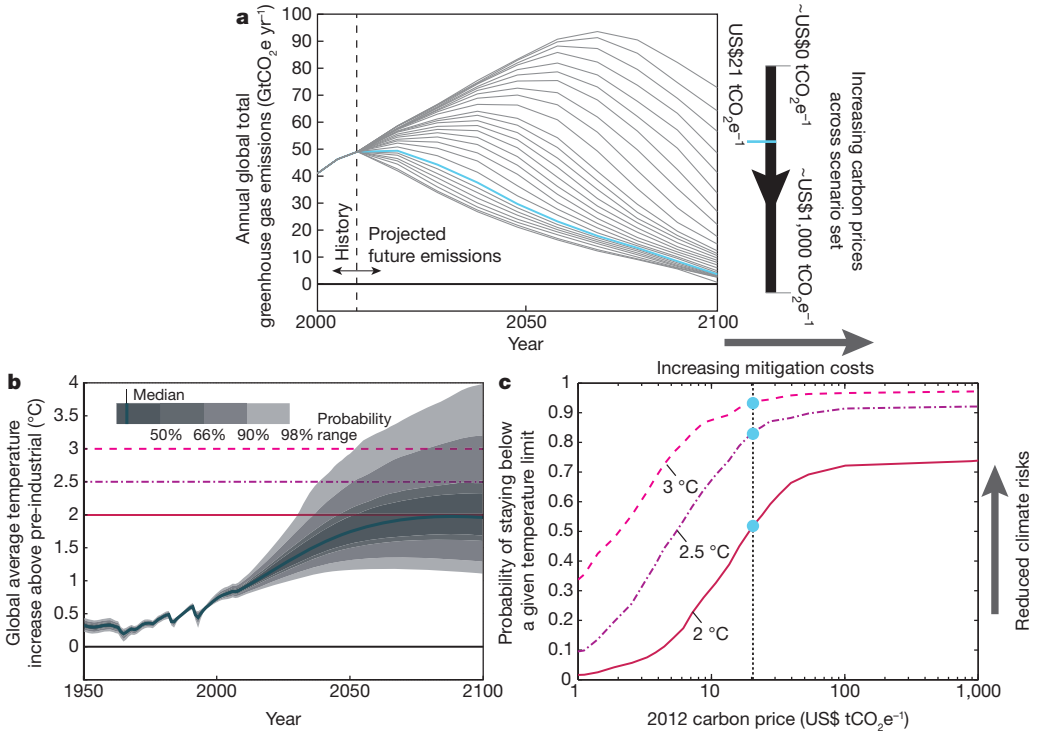


Figure 1 | Methodology for creating cost–risk relationships for a given temperature limit. **a**, Illustrative set of emissions scenarios with carbon prices increasing from zero to US$1,000 tCO_2e^{-1}. The arrow at right indicates the direction of increasing carbon prices across the illustrative set. The blue-highlighted scenario has a global carbon price discounted back to 2012 of US$21 tCO_2e^{-1}. **b**, Probabilistic temperature projections for the blue trajectory in **a**. Horizontal lines at 2, 2.5 and 3 °C show possible target temperature limits. In this illustrative scenario, median (50% probability) warming is 2.0 °C. There is a slim chance (<5%) that temperatures remain below 1.3 °C and a large chance (>90%) that they remain below 3.0 °C. **c**, Cumulative distributions of carbon prices consistent with limiting warming to below 2, 2.5 and 3 °C, as indicated. Blue dots indicate points defined by the cost information of the scenario highlighted in **a** and the probabilistic temperature projection in **b**.

enough (Fig. 2a). In certain instances, however, the lack of mitigation options (such as renewable technologies, nuclear power or limited biomass and afforestation potential) could require substantially higher carbon prices to keep this target viable. At the limit is CCS: the complete elimination of this mitigation option—either for technological reasons or as a result of social and political concerns—would put the 2 °C objective (with more than 66% probability) out of reach in our model, no matter how high the carbon price.

The future availability of energy supply technologies (for example renewable technologies and CCS) tells only one side of the story; a strong finding from our analysis is that social developments influencing energy demand (that is, efficiency of energy use) are even more important. This is evidenced in Fig. 2b by the differences between three distinct scenario families whose future energy demands vary greatly (low, intermediate and high; see Table 1 and ref. 7 for details). In the low-demand scenarios, end-use efficiency measures and conservation-minded energy and urban planning policies are instituted ubiquitously throughout the industrial, building and transportation sectors in all countries. This leads to a global energy demand in 2050 that is about 25% lower than our intermediate baseline, which broadly applies historical patterns of efficiency improvement[7]. Such reductions in demand could be crucial in keeping the 2 °C objective within reach, independently of what happens in terms of energy supply.

For example, in our scenarios the availability of nuclear power has an almost negligible effect on overall mitigation costs compared with a switch from a scenario with an intermediate energy demand to one with a high demand. Low-demand strategies would ensure a higher likelihood of staying below 2 °C for the same carbon price (from 66% to more than 80% at US$40 tCO_2e^{-1}), or, viewed in a different way, would drastically reduce the cost of reaching the 'likely'[19] probability level (from US$40 tCO_2e^{-1} to around US$10–15 tCO_2e^{-1}). In contrast, a high-energy-demand future—about 20% greater in 2050 than the intermediate baseline, resulting from more energy-intensive lifestyles and less efficiency- and conservation-focused policies—would require much higher carbon prices (>US$150 tCO_2e^{-1}) and make it much more difficult, if not impossible, to reach the 2 °C objective with a probability of more than 66%.

Overall, Fig. 2b indicates that the present influence of geophysical uncertainties on the spread in mitigation costs to achieve the 2 °C objective is comparable to that of the uncertainties arising from different future pathways for social development and technological changes and choices. The maximum difference in probability of staying below 2 °C between the least costly (blue-dashed) and the most costly (red-dotted) distribution is slightly greater than 60 percentage points. This roughly matches the range of probabilities found when taking into account the Earth system uncertainty under the same supply and demand assumptions (for example 0–70% for the entire range of carbon prices in our central case, which assumes a reference technology portfolio and intermediate energy demand). Such a finding is broadly consistent with earlier studies comparing the relative contributions of geophysical and technological factors[22] using a non-probabilistic approach.

Yet, despite all of the uncertainty in the geophysical, social and technological aspects, our analysis indicates that the dominant factor affecting the likelihood and costs of achieving the 2 °C objective is politics. Here we model political uncertainties by varying the timing of concerted global mitigation efforts. Although studies of the implication of delays in climate action are not new[2,23–26], our results show how geophysical uncertainties interact and compare with political inertia: if global temperature rise is to be kept below 2 °C with a probability of more than 66% under central technology and energy demand assumptions, our scenarios show that immediate and globally coordinated mitigation action is necessary (Fig. 2c; Supplementary Information provides an explanation of 'immediate'). Only for low-energy-demand pathways can global mitigation action be delayed until 2020 and the

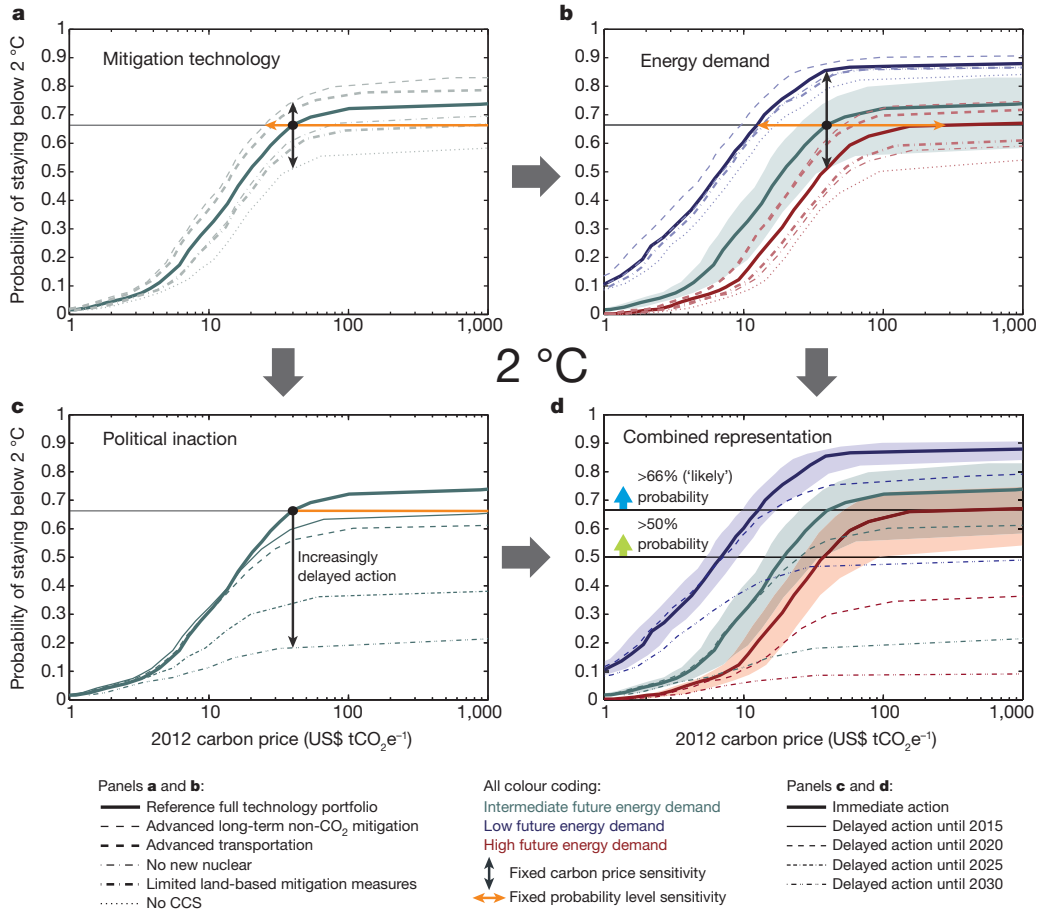


Figure 2 | Influence of mitigation technology, energy demand and political inaction on the cost–risk distributions for staying below 2 °C. Cost distributions for six cases with varying future availability of specific mitigation technologies (**a**) and three sensitivity cases for future energy demand (**b**, thick solid lines). Shaded areas and dashed lines in **b** represent technology-sensitivity cases comparable to those shown in **a**. Shaded areas and dashed lines in **d** represent technology- and politics-sensitivity cases comparable to those in **b** and **c**, respectively. **c**, Impact of delayed global mitigation action. **d**, Overview figure combining all sensitivity cases. The horizontal line in **a**–**c** is the 66% line. Similar figures for 2.5 and 3 °C are provided in Supplementary Figs 4 and 5. A comparison with 91 scenarios from the literature[2] is provided in Supplementary Fig. 7.

2 °C objective still be achieved with a probability of more than 66% (or delayed until 2030 with a probability of 50%; Fig. 2d).

In conclusion, we find that the effect of global mitigation action delayed by two decades is much more pronounced than the consequences of uncertainty surrounding mitigation technology availability and future energy demands, and renders even the geophysical uncertainties almost irrelevant for the 2 °C objective (Fig. 2d, Supplementary Table 2 and Supplementary Fig. 9). Furthermore, we find asymptotic limits to increasing the probability of reaching a given temperature objective in our model: if mitigation action is delayed, simply spending more money on the problem in the future will not increase this probability beyond certain limits imposed by the Earth system.

Our mitigation cost distribution methodology can also be applied to other temperature objectives, for example a weaker limit (3 °C) or a

Table 1 | Overview of sensitivity cases

Mitigation technology	
Technological limits	
No new nuclear	From 2020 onwards, no new investments are made into nuclear power, leading to a full phase-out of existing plants by 2060.
Limited land-based measures	The mitigation potential from biomass, land use and forestry is limited.
No CCS	Technology to capture and geologically store CO_2 (CCS) from fossil fuel and/or biomass energy never becomes available on a globally meaningful scale.
Technological breakthroughs	
Advanced transportation	Fundamental changes in transportation infrastructures (for example for electric transport) or major breakthroughs in transportation technology (for example in hydrogen fuel cells) lead to increased decarbonization of the transportation sector.
Advanced non-CO_2 mitigation	The mitigation potential of non-CO_2 greenhouse gases is assumed to improve continuously, beyond the level of current best practice.
Energy demand	
Intermediate demand	The development of energy demand and efficiency improvements is broadly consistent with (only slightly faster than) what is observed historically.
High demand	Energy efficiency improves more slowly than historically observed, leading to a high future energy demand.
Low demand	Energy efficiency improves radically in all end-use sectors (buildings, industry, transport) leading to low future energy demand.
Political inaction	
Delayed action	Globally concerted mitigation action is postponed from today until 2015, 2020, 2025 and 2030 in respective cases.

Detailed descriptions and background are provided in Supplementary Information and ref. 7.

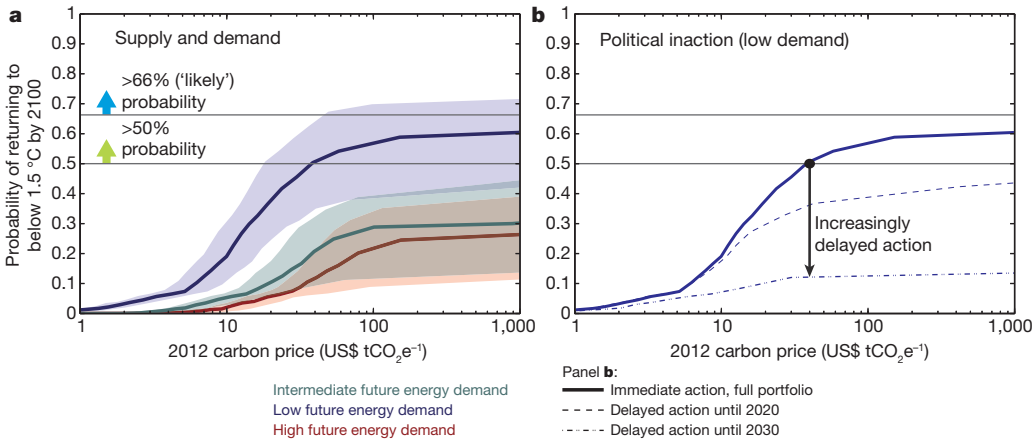


Figure 3 | Cost–risk distributions for returning global temperature increase to below 1.5 °C by 2100. a, Cost distributions for three cases of varying future energy demand (solid red, green, and blue lines) and varying future availability of specific mitigation technologies (shaded ranges around solid lines). Ranges show the variation over all assessed technology-sensitivity cases. **b**, Influence of global mitigation action delayed from now until 2030. The vertical axis shows the probability of returning global average temperature increase to below 1.5 °C by 2100.

stricter limit (1.5 °C), the second of which has already been discussed in the policy arena[21]. We find that unless energy demand is low, CCS technology is available and global climate action is undertaken immediately, holding temperature increase to below 1.5 °C by 2100 with a probability of at least 50% is already unfeasible (Fig. 3a). In terms of costs, this would require the immediate introduction of global carbon prices of more than US$40 tCO_2e^{-1} (increasing over time with the discount rate). If global mitigation action were delayed by 10 to 20 years, a carbon price of US$40 tCO_2e^{-1} would yield probabilities of only 10–35%; and, even under higher prices, a 50% probability could no longer be reached under central technology and low-energy-demand assumptions (Fig. 3b). However, the same carbon price, US$40 tCO_2e^{-1}, would prevent an increase in warming beyond 3 °C with a high probability (>90%) for all supply–demand combinations, contingent on the immediate global introduction of the pricing instrument (Supplementary Fig. 5).

Our findings have implications for the ongoing international climate policy discussions[27], which foresee a global agreement coming into effect only in 2020. For this delay strategy to be successful, national and local governments would need to place far greater importance on concurrent demand-side solutions to climate protection (thus lowering energy demand growth), as well as on voluntary or revised near-term mitigation policies and measures that anticipate and are consistent with a future stringent climate agreement. Our model results show that robustly safeguarding the future achievement of the oft-discussed 2 °C objective requires that society embarks on a higher-efficiency, lower-energy-demand course well before 2020 in the context of sustained, concerted and coordinated mitigation efforts.

METHODS SUMMARY

We create a large ensemble ($n > 700$) of emissions scenarios with MESSAGE[14,15], a global integrated assessment modelling framework with a detailed representation of greenhouse-gas-emitting sectors, by imposing cumulative constraints on greenhouse gas emissions (for all such gases: carbon dioxide, methane, nitrous oxide, halocarbons and fluorinated gases) of varying stringencies for the whole twenty-first century, and by changing salient features in the underlying scenario assumptions (see Supplementary Information, Supplementary Fig. 1, Supplementary Table 1 and ref. 7 for a full set of assumptions). Our scenarios assume 'middle-of-the-road' assumptions for socio-economic development from previous research on scenarios: population peaking at 9,700,000,000 later this century (United Nations median projection[28]) and gross world product increasing more than sevenfold by 2100 (updated Special Report on Emissions Scenarios B2 scenario projection by the Intergovernmental Panel on Climate Change[7,29]).

We then compute probabilistic estimates of global temperature increase for each scenario with the MAGICC climate model[16,17,30]. These estimates are based on a 600-member ensemble of temperature projections for each scenario, which together closely represent the carbon-cycle and climate uncertainties as assessed in the Fourth Assessment Report of the Intergovernmental Panel on Climate Change[17]. Additionally, our temperature projections are also constrained by observations and estimates of hemispheric temperatures and ocean heat uptake (Supplementary Information). The probability of staying below a given temperature threshold is computed over the entire twenty-first century and relative to pre-industrial levels. In contrast to the 2 °C objective, the target of 1.5 °C is referred to as a long-term goal[21], meaning that we allow a small, temporary overshoot and assess the probability of returning warming to below 1.5 °C by 2100.

We present our results using carbon prices as the cost metric. For an illustration of our results using other cost metrics, such as total mitigation costs, see section 1.4 of Supplementary Information and Supplementary Figs 2 and 3. The carbon price shown is the price at the time action starts, discounted back to 2012 with a discount rate of 5% per year (Supplementary Information, section 1.4, and Supplementary Fig. 6).

Received 15 August; accepted 5 November 2012.

1. Randalls, S. History of the 2 °C climate target. *Clim. Change* **1,** 598–605 (2010).
2. Clarke, L. *et al.* International climate policy architectures: overview of the EMF 22 International Scenarios. *Energy Econ.* **31,** S64–S81 (2009).
3. Edenhofer, O. *et al.* The economics of low stabilization: model comparison of mitigation strategies and costs. *Energy J.* **31,** 11–48 (2010).
4. UNEP. *Bridging the Emissions Gap* 15–20 (United Nations Environment Programme, 2011).
5. O'Neill, B. C., Riahi, K. & Keppo, I. Mitigation implications of midcentury targets that preserve long-term climate policy options. *Proc. Natl Acad. Sci. USA* **107,** 1011–1016 (2010).
6. Core Writing Team, Pachauri, R. K. & Reisinger, A. (eds) *Climate Change 2007: Synthesis Report* (Intergovernmental Panel on Climate Change, 2007).
7. Riahi, K. *et al.* in *Global Energy Assessment: Toward a Sustainable Future* 1203–1306 (Cambridge Univ. Press, 2012).
8. Friedlingstein, P. *et al.* Climate–carbon cycle feedback analysis: results from the C4MIP model intercomparison. *J. Clim.* **19,** 3337–3353 (2006).
9. Meehl, G. A., Covey, C., McAvaney, B., Latif, M. & Stouffer, R. J. Overview of the coupled model intercomparison project. *Bull. Am. Meteorol. Soc.* **86,** 89–93 (2005).
10. Meehl, G. A. *et al.* in *IPCC Fourth Assessment Report* (eds S. Solomon *et al.*) 747–847 (Cambridge Univ. Press, 2007).
11. Meinshausen, M., Wigley, T. M. L. & Raper, S. C. B. Emulating atmosphere-ocean and carbon cycle models with a simpler model, MAGICC6 – Part 2: applications. *Atmos. Chem. Phys.* **11,** 1457–1471 (2011).
12. Archer, D. *et al.* Atmospheric lifetime of fossil fuel carbon dioxide. *Annu. Rev. Earth Planet. Sci.* **37,** 117–134 (2009).
13. Schaeffer, M., Kram, T., Meinshausen, M., van Vuuren, D. P. & Hare, W. L. Near-linear cost increase to reduce climate-change risk. *Proc. Natl Acad. Sci. USA* **105,** 20621–20626 (2008).
14. Rao, S. & Riahi, K. The role of non-CO_2 greenhouse gases in climate change mitigation: long-term scenarios for the 21st century. *Energy J.* **27,** 177–200 (2006).
15. Riahi, K., Gruebler, A. & Nakicenovic, N. Scenarios of long-term socio-economic and environmental development under climate stabilization. *Technol. Forecast. Social Change* **74,** 887–935 (2007).
16. Meinshausen, M., Raper, S. C. B. & Wigley, T. M. L. Emulating coupled atmosphere-ocean and carbon cycle models with a simpler model, MAGICC6 – Part 1: Model description and calibration. *Atmos. Chem. Phys.* **11,** 1417–1456 (2011).

17. Rogelj, J., Meinshausen, M. & Knutti, R. Global warming under old and new scenarios using IPCC climate sensitivity range estimates. *Nature Clim. Change* **2,** 248–253 (2012).
18. Knutti, R. *et al.* A review of uncertainties in global temperature projections over the twenty-first century. *J. Clim.* **21,** 2651–2663 (2008).
19. Mastrandrea, M. D. *et al. Guidance Note for Lead Authors of the IPCC Fifth Assessment Report on Consistent Treatment of Uncertainties* 3 (Intergovernmental Panel on Climate Change, 2010); available at http://www.ipcc.ch/pdf/supporting-material/uncertainty-guidance-note.pdf.
20. Smith, S. M. *et al.* Equivalence of greenhouse-gas emissions for peak temperature limits. *Nature Clim. Change* **2,** 535–538 (2012).
21. UNFCCC. *FCCC/CP/2010/7/Add.1 Decision 1/CP.16* 3 (UN Framework Convention on Climate Change, 2010).
22. Smith, S. J. & Edmonds, J. A. The economic implications of carbon cycle uncertainty. *Tellus B* **58,** 586–590 (2006).
23. Vaughan, N., Lenton, T., &. Shepherd, J. Climate change mitigation: trade-offs between delay and strength of action required. *Clim. Change* **96,** 29–43 (2009).
24. den Elzen, M., van Vuuren, D. & van Vliet, J. Postponing emission reductions from 2020 to 2030 increases climate risks and long-term costs. *Clim. Change* **99,** 313–320 (2010).
25. Bosetti, V., Carraro, C., Sgobbi, A. & Tavoni, M. Delayed action and uncertain stabilisation targets. How much will the delay cost? *Clim. Change* **96,** 299–312 (2009).
26. Krey, V. & Riahi, K. Implications of delayed participation and technology failure for the feasibility, costs, and likelihood of staying below temperature targets: greenhouse gas mitigation scenarios for the 21st century. *Energy Econ.* **31,** S94–S106 (2009).
27. UNFCCC. *FCCC/CP/2011/9/Add.1 Decision 1/CP.17* (UN Framework Convention on Climate Change, 2011).
28. UN. *World Population Prospects: The 2008 Revision Population Database* (United Nations, 2009).
29. Nakicenovic, N. & Swart, R. *IPCC Special Report on Emissions Scenarios* (Cambridge Univ. Press, 2000).
30. Meinshausen, M. *et al.* Greenhouse-gas emission targets for limiting global warming to 2°C. *Nature* **458,** 1158–1162 (2009).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank V. Krey, P. Kolp and M. Strubegger for their support in developing the model set-up and extracting the results, R. Knutti and R. Socolow for comments and feedback during the writing process and S. Hatfield-Dodds, whose review comments substantially contributed to improving our manuscript. J.R. was supported by the Swiss National Science Foundation (project 200021-135067) and the IIASA Peccei Award Grant.

Author Contributions All authors were involved in designing the research. J.R. performed the research in collaboration with D.L.M. All authors contributed to writing the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.R. (joeri.rogelj@env.ethz.ch).

LETTER

doi:10.1038/nature11679

The oxidation state of the mantle and the extraction of carbon from Earth's interior

Vincenzo Stagno[1]†, Dickson O. Ojwang[1], Catherine A. McCammon[1] & Daniel J. Frost[1]

Determining the oxygen fugacity of Earth's silicate mantle is of prime importance because it affects the speciation and mobility of volatile elements in the interior and has controlled the character of degassing species from the Earth since the planet's formation[1]. Oxygen fugacities recorded by garnet-bearing peridotite xenoliths from Archaean lithosphere are of particular interest, because they provide constraints on the nature of volatile-bearing metasomatic fluids and melts active in the oldest mantle samples, including those in which diamonds are found[2,3]. Here we report the results of experiments to test garnet oxythermobarometry equilibria[4,5] under high-pressure conditions relevant to the deepest mantle xenoliths. We present a formulation for the most successful equilibrium and use it to determine an accurate picture of the oxygen fugacity through cratonic lithosphere. The oxygen fugacity of the deepest rocks is found to be at least one order of magnitude more oxidized than previously estimated. At depths where diamonds can form, the oxygen fugacity is not compatible with the stability of either carbonate- or methane-rich liquid but is instead compatible with a metasomatic liquid poor in carbonate and dominated by either water or silicate melt. The equilibrium also indicates that the relative oxygen fugacity of garnet-bearing rocks will increase with decreasing depth during adiabatic decompression. This implies that carbon in the asthenospheric mantle will be hosted as graphite or diamond but will be oxidized to produce carbonate melt through the reduction of Fe^{3+} in silicate minerals during upwelling. The depth of carbonate melt formation will depend on the ratio of Fe^{3+} to total iron in the bulk rock. This 'redox melting' relationship has important implications for the onset of geophysically detectable incipient melting and for the extraction of carbon dioxide from the mantle through decompressive melting.

Oxythermobarometry measurements[4,5] indicate that the oxygen fugacity (f_{O_2}) recorded by spinel-bearing peridotite rocks, which originate at a depth of 30–50 km in the mantle, ranges from approximately 3 log units below to 2 log units above the fayalite–magnetite–quartz oxygen buffer (that is, the $\log(f_{O_2})$ value determined relative to the buffering equilibrium $3Fe_2SiO_4 + O_2 = 2Fe_3O_4 + 3SiO_2$; we denote this value ΔFMQ). Abyssal peridotites, thought to be residues of melting to produce mid-ocean-ridge basalt (MORB), lie at the low end of this range, and xenoliths associated with subduction zones are in general at the higher end. Over this range of f_{O_2}, CO_2 and H_2O are the major volatile species, consistent with their dominance in erupting basaltic magmas[5,6].

For garnet-bearing rocks, which originate at depths between 50 and 220 km, the only commonly employed oxythermobarometry equilibrium uses the $Fe_3Fe_2^{3+}Si_3O_{12}$ (skiagite) garnet component:

$$\underset{\text{garnet}}{2Fe_3Fe_2^{3+}Si_3O_{12}} = \underset{\text{olivine}}{4Fe_2SiO_4} + \underset{\text{orthopyroxene}}{2FeSiO_3} + O_2 \quad (1)$$

This equilibrium, which requires the precise determination of the ratio of Fe^{3+} to total Fe ($Fe^{3+}/\sum Fe$) in garnet, has been extensively used to determine the redox state of xenoliths mainly from cratonic lithosphere, but has only been experimentally tested for its accuracy in estimating f_{O_2} under conditions of 3 and 3.5 GPa and 1,300 °C, equivalent to ~100-km depth[7]. The conclusion from the use of equilibrium (1) has been that the f_{O_2} value of the Archaean lithosphere decreases with depth until conditions are close to those at which Fe-Ni metallic alloy would start to precipitate from mantle minerals[3,8] (5 log units below the FMQ buffer). Under such conditions, volatile species in the mantle would be dominated by CH_4 and H_2 rather than by H_2O and CO_2 (refs 8–10).

To test oxythermobarometry equilibria, we performed multi-anvil experiments at high pressure and temperature on garnet peridotite assemblages and determined the $Fe^{3+}/\sum Fe$ ratio of garnet at f_{O_2} values that we measured using an Ir-Fe alloy redox sensor[11]. Experimental capsules comprised layers of garnet powder (either natural or formed from an initial synthetic glass) sandwiched between layers containing a mixture of olivine, orthopyroxene and garnet. In most experiments, the oxygen fugacity was buffered by a mixture of graphite and carbonate mixed within the sample, although several experiments contained H_2O rather than carbonate and one experiment involved no carbon. Approximately 3% Ir was added to all layers of the experiments. This alloys with Fe from the silicates during the experiment, allowing f_{O_2} to be determined from the equilibrium

$$\underset{\text{olivine}}{Fe_2SiO_4 + Mg_2SiO_4} = \underset{\text{orthopyroxene}}{2MgSiO_3} + \underset{\text{alloy}}{2Fe} + O_2 \quad (2)$$

The Fe content of the Ir-Fe f_{O_2}-sensor shifts to equilibrate with f_{O_2}, which can be measured by chemical analysis of the silicate and alloy phases. The ferric Fe contents of the garnet layers in the run products were determined by Mössbauer spectroscopy. Experiments were conducted at pressures of 3, 6 and 7 GPa (corresponding to depths of 90–250 km) and temperatures of 1,300–1,600 °C for periods of 12–24 h.

The resulting garnet Fe^{3+} contents (~5–18% of total Fe) are similar to values found in garnets from mantle xenoliths[2]. As can be seen in Fig. 1, oxygen fugacities calculated using equilibrium (1) are in good agreement with those determined from the Ir-Fe sensor equilibrium (2) at 3 GPa but deviate strongly by 1–2 log units at 6 and 7 GPa. The origin of this discrepancy at higher pressure is unclear, although we note that constraints on thermodynamic data for the $Fe_3Fe_2^{3+}Si_3O_{12}$ garnet component skiagite are sparse[7,12] (Methods). A further oxythermobarometer equilibrium

$$\underset{\text{garnet}}{2Ca_3Fe_2Si_3O_{12}} + \underset{\text{garnet}}{2Mg_3Al_2Si_3O_{12}} + \underset{\text{orthopyroxene}}{4FeSiO_3} = \underset{\text{garnet}}{2Ca_3Al_2Si_3O_{12}} + \underset{\text{olivine}}{4Fe_2SiO_4} + \underset{\text{orthopyroxene}}{6MgSiO_3} + O_2 \quad (3)$$

was proposed in a previous study[2] and has been successfully tested at 3 GPa and 1,300 °C (ref. 7). The equilibrium is based on the ferric-Fe-bearing garnet component andradite ($Ca_3Fe_2Si_3O_{12}$), for which thermodynamic data are well constrained[13], and although thermodynamic data for the $Fe_3Fe_2^{3+}Si_3O_{12}$ garnet component are still required

[1]Bayerisches Geoinstitut, University of Bayreuth, Bayreuth D-95440, Germany. †Present address: Geophysical Laboratory, Carnegie Institution of Washington, Washington DC 20015, USA.

for treatment of the reciprocal reactions, these are likely to be less sensitive to errors in pressure and temperature dependence (Methods). As shown in Fig. 1, oxygen fugacities calculated from the experimental data using equilibrium (3) are in much better agreement with those determined from the Ir-Fe sensor over the entire pressure range when compared with the results for equilibrium (1).

In Fig. 2, oxygen fugacities calculated using equilibrium (3) for xenoliths from the Kaapvaal, Slave and Siberian cratons are shown as functions of equilibration depth[2,3,14–19]. The deepest determined peridotite oxygen fugacities are at least 1 log unit more oxidized than previous estimates and extend to only slightly more-reducing conditions than do abyssal peridotites that are considered to be MORB melting residues. The most reduced samples ($<\Delta$FMQ -3) correlate well with indicators for extreme melt depletion such as high Cr contents and high olivine Mg/(Fe+Mg) ratio, and more oxidized samples probably result from subsequent metasomatism by melt or fluid phases. In the diamond stability field, the range of xenolith oxygen fugacities is quite narrow and lies between ΔFMQ -2 and ΔFMQ -3.5. This range is below that defined by the reaction

$$\underset{\text{enstatite}}{MgSiO_3} + \underset{\text{magnesite}}{MgCO_3} = \underset{\text{olivine}}{Mg_2SiO_4} + \underset{\text{diamond}}{C} + O_2 \quad (4)$$

which marks the upper f_{O_2} stability limit of diamond with respect to the solid carbonate mineral magnesite[20] and is denoted EMOD. Because temperatures in the lithosphere can be above the carbonated peridotite solidus[21], an equilibrium equivalent to equilibrium (4) determines the stability of diamond with respect to pure carbonate melt[22] (Fig. 2, dashed curve; $MgCO_3$ melt). Further curves show the stability of diluted carbonate melts with respect to diamond in terms of the molar percentage of the carbonate (CO_3^{2-}) component. The majority of xenoliths from the diamond stability field have oxygen fugacities that would not be in equilibrium with pure carbonate melts but only with melt phases in which the carbonate component (or CO_2 component in a fluid) is relatively dilute, that is, 1–10%. For example, the diluting component could be H_2O or, if temperatures were sufficiently high, silicate melt. The latter would be more consistent with low mineral H_2O contents recently measured for such samples[23], implying metasomatism at temperatures substantially higher than the present average mantle temperature. However, the range of f_{O_2} is above that where fluid or melt phases would contain a considerable CH_4 component. The results, therefore, rule out CH_4- and $MgCO_3$-rich fluids or melts being major pervasive metasomatic agents in the Archaean lithosphere.

Equilibrium (3) will also exert some control over the oxygen fugacity of adiabatically upwelling mantle, which can influence the inception of decompression melting beneath mid-ocean ridges. In Fig. 3, the oxygen fugacity of a garnet peridotite with a bulk silicate Earth composition[24] is calculated as a function of the bulk rock $Fe^{3+}/\sum Fe$ ratio using equilibrium (3) (Methods Summary). Experimental partitioning data, and data on the Fe^{3+} distribution in xenoliths, are used for this calculation (Methods). As pressure decreases, the oxygen fugacity for a given bulk rock $Fe^{3+}/\sum Fe$ ratio will increase, causing a packet of mantle to auto-oxidize at a rate of ~1/60 log units per kilometre[9]. Figure 3 also shows a curve indicating the stability limit of diamond and graphite with respect to carbonate melt, determined along an adiabat. At oxygen fugacities below this curve diamond and graphite are stable, whereas above it carbonate melts can form. This curve dips at low pressure owing to the onset of the main phase of silicate melting[22].

Estimates for average mantle $Fe^{3+}/\sum Fe$ content are of the order of 2%. However, this may be a minimum estimate because it is based on analyses of lithospheric mantle samples[10,14]; Fig. 3 indicates that Fe^{3+} contents of 3–4% of total Fe yield oxygen fugacities closer to those determined for the MORB source from basalt glass analyses[25]. At depths of 200 km, carbon in mantle containing less than 4% Fe^{3+} would be in the diamond stability field and only when decompressed to 150 km would the oxygen fugacity be compatible with the oxidation of graphite to form carbonate melt through the reduction of Fe^{3+} in

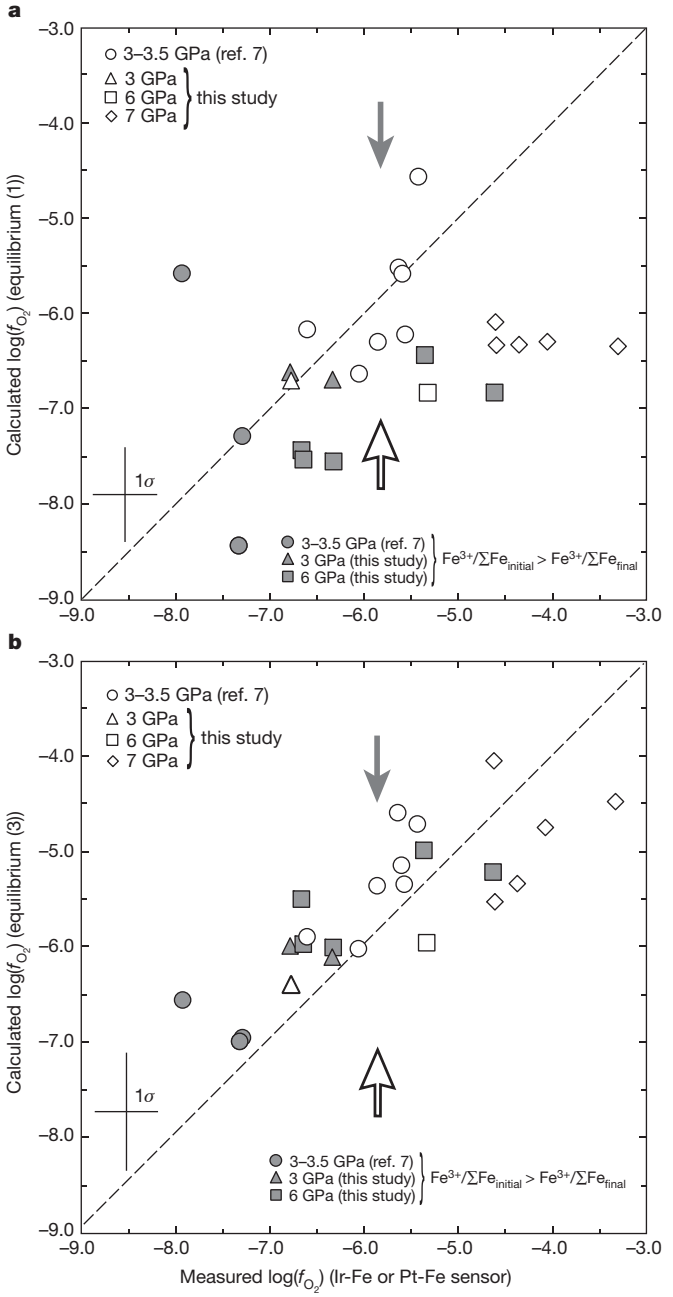


Figure 1 | Comparison between oxythermobarometers. **a**, Oxygen fugacities calculated using the Ir-Fe redox sensor, equilibrium (2), for garnet–peridotite assemblages equilibrated in this study are compared with those calculated using equilibrium (1) and the model from ref. 7. **b**, The f_{O_2} estimates from the Ir-Fe redox sensor compared with those calculated using the model proposed in this study for equilibrium (3). For comparison we show data at 3 and 3.5 GPa from ref. 7, where f_{O_2} was calculated using a similar Pt-Fe redox sensor. Reversal experiments are shown using filled symbols and indicate experiments that used synthetic and natural garnets with initial $Fe^{3+}/\sum Fe$ ratios that were greater than those measured in the recovered garnet samples, whereas open symbols indicate lower initial garnet $Fe^{3+}/\sum Fe$ ratios. The arrows indicate the directions from which the corresponding data approach equilibrium. The oxygen fugacities calculated using equilibrium (1) deviate considerably from the alloy redox sensor values at pressures of 6–7 GPa, whereas those calculated with equilibrium (3) are in good agreement with the alloy sensor over the entire pressure range. Uncertainties (1σ) in the Ir-Fe and Pt-Fe log(f_{O_2}) measurements are 0.3 log units, and those estimated for equilibria (1) and (3) are 0.5 and 0.7 log units, respectively (Methods).

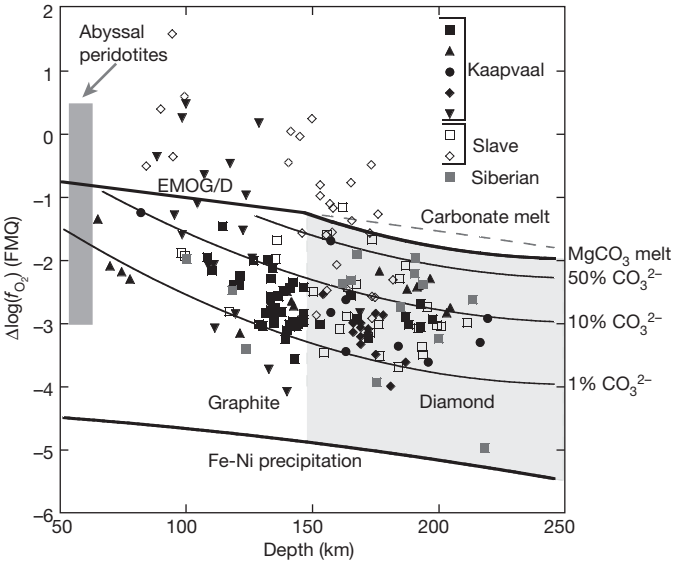


Figure 2 | Log(f_{O_2}) (normalized to the FMQ buffer) calculated for xenoliths from the cratonic lithosphere using equilibrium (3). In the key, black symbols indicate mantle xenoliths from the Kaapvaal craton (up-triangles, ref. 2; squares, ref. 3; circles, ref. 14; diamonds, ref. 16; down-triangles, ref. 17), open symbols represent the Slave craton (squares, ref. 15; diamond, ref. 18) and grey squares represent the Siberian craton[19]. The oxygen fugacity of the Fe-Ni precipitation curve indicates conditions under which nickel-rich iron metal will start to precipitate from coexisting silicates[8–10]. The curve labelled EMOG/D is the oxygen fugacity calculated for equilibrium (4) along a cratonic geotherm. It defines the stability field between diamond (or graphite) and magnesite ($MgCO_3$) in the mantle. At depths >150 km, this curve dips owing to carbonate melting. The curves plotted below EMOG/D indicate oxygen fugacities for melts in equilibrium with diamond or graphite, for different carbonate (CO_3^{2-}) contents[22] (molar percentage of MCO_3 is indicated for each curve, where M is a divalent cation). The majority of cratonic xenoliths would coexist with melts characterized by dilute carbonate contents (1–10%). The light-grey region indicates the diamond stability field. The dark-grey rectangle indicates the range of oxygen fugacities exhibited by abyssal peridotites[5].

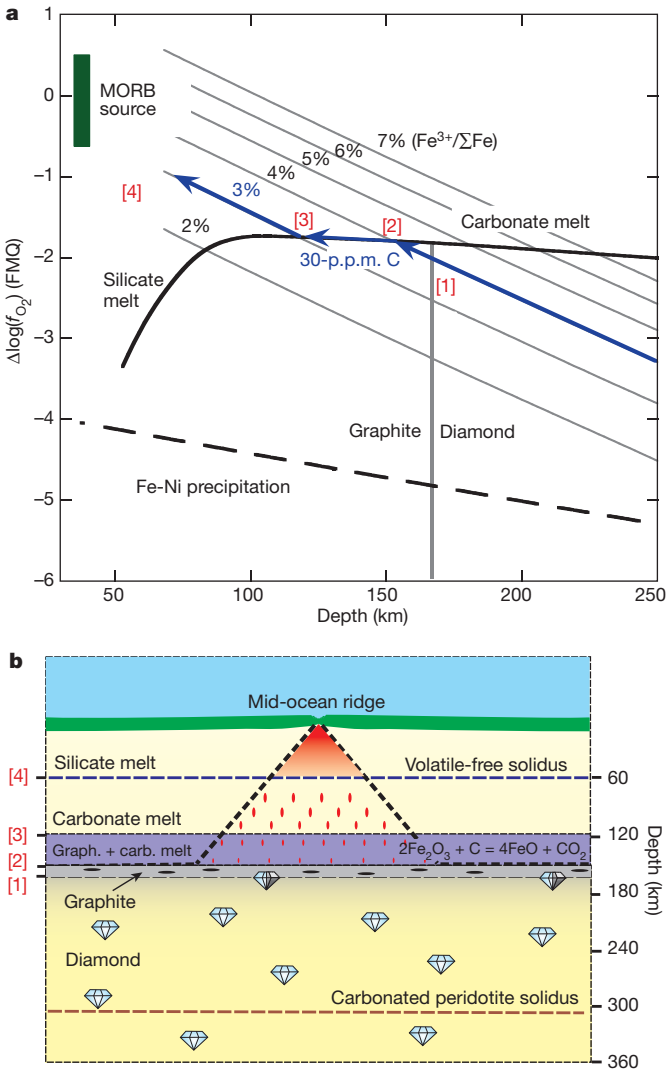


Figure 3 | Speciation of carbon in adiabatically upwelling mantle. **a**, Estimates for the oxygen fugacity of a bulk silicate Earth mantle[24], determined for different bulk rock $Fe^{3+}/\sum Fe$ ratios calculated along a 1,320 °C adiabat are indicated by the parallel grey curves. These curves are determined from equation (6) using experimental and natural mineral partitioning data[2,14,34,35]. The bold black curve indicates the oxygen fugacity of an assemblage buffered by graphite (or diamond) coexisting with carbonate melt[22] and dips at low pressure owing to the onset of silicate melting. The green rectangle indicates the range of f_{O_2} estimated for the MORB source[25]. Mantle containing 4% $Fe^{3+}/\sum Fe$ would be in the diamond stability field at depths >170 km. At shallower depths, transformation to graphite would occur at point [1]. At 150 ± 50 km, the oxygen fugacity of upwelling mantle would intersect the equilibrium between graphite and carbonate melt (point [2]), which would initiate carbonate melting through the reduction of Fe^{3+} in silicate phases. This 'redox melting' process would occur over a depth interval of ~30 km, between points [2] and [3], over which 30 p.p.m. of carbon in the mantle source would be oxidized by the reduction of the bulk rock $Fe^{3+}/\sum Fe$ ratio by 1%. For carbon contents >30 p.p.m. in the mantle source, a greater initial $Fe^{3+}/\sum Fe$ ratio is required for the oxygen fugacity to remain in the range estimated for the MORB source, which is remarkably uniform[25]. At point [4], the carbonate melt will evolve towards a silicate melt composition. **b**, The path followed by the blue arrows in **a** is demonstrated with the same points indicated. Carbonate melts formed at depth as a result of redox melting are potentially focused into the source region of basalts[31] (dashed diagonal lines). Redox melting to produce carbonate melt occurs at a depth ~150 km shallower than the carbonated peridotite solidus.

silicates[6]. The uncertainty in the depth of initial carbon oxidation is ~50 km once the errors in the determinations of f_{O_2} are taken into account. The partitioning of water from mineral phases into the melt may increase this depth by a similar magnitude. The reduction of Fe^{3+} would force the oxygen fugacity to remain buffered[26] along the equilibrium curve between graphite and carbonate melt until all graphite in the rock had been oxidized to carbonate. The oxidation of 30-p.p.m. graphite, which is a typical carbon content estimate of the MORB source[21,27], would occur over a depth interval of approximately 30 km and only then would auto-oxidation continue to increase the oxygen fugacity.

Typical estimated MORB source carbon contents are sufficient to reduce the bulk Fe^{3+} content of the mantle by 1% of total Fe during decompression, implying that the mantle at depth is richer in Fe^{3+} by an amount proportional to the ultimate CO_2 content of the MORB source. This has several consequences. First, geophysical studies have been used to argue for the presence of incipient carbonate melting at depths >200 km beneath mid-ocean ridges[28,29]. If this were the case then oxidation of 30-p.p.m. carbon to carbonate at this depth would require the bulk mantle Fe^{3+} content to be >7 ± 2% of total Fe, which is higher than previous estimates[10,14]. Second, although estimates for MORB mantle C contents are 30 p.p.m., estimates for some enriched sources—based, for example, on CO_2 contents of ocean island basalts—have C contents in the range 300–1,300 p.p.m. (ref. 21). Several studies have indicated that the formation of majoritic garnet additionally contributes to a decrease in mantle f_{O_2} such that the mantle probably contains Fe-Ni-rich metallic alloy at depths >300 km (refs 9, 10, 30). In the transition zone, and the lower mantle in particular, the stability of minerals with substantial Fe^{3+} contents at low oxygen fugacity would ensure that even large mantle CO_2 contents would reduce to form diamond, CH_4, Fe-Ni carbide or C-bearing metallic

alloy[8–10]. To convert even this lowest estimate of 300-p.p.m. C into CO_2 on decompression would require the ocean island basalt source at depth to have >10% of total Fe as Fe^{3+}.

This large Fe^{3+} content might imply the existence of either an Fe^{3+}-rich primordial mantle source[8] or the involvement of recycled oceanic lithosphere richer in both C and Fe^{3+}, possibly as a result of the pressure-induced reduction of carbonates[31]. However, C contents of ocean island basalt sources might be overestimated as a result of carbonate-melt focusing. The inception of deep carbonate melting during upwelling, due to depression of the carbonated-peridotite solidus, can potentially focus carbonate-rich melts extracted from much larger volumes of the mantle into the smaller source volume from which the bulk of basaltic magma is extracted[31,32]. Deep carbonate-rich melting, however, requires higher initial mantle Fe^{3+} contents, which couples the extraction of carbon from the mantle to the oxygen fugacity. High CO_2 contents of basaltic magmas, therefore, do not necessarily imply high carbon source concentrations but instead plausibly imply more-oxidized sources. A more-reduced mantle in Earth's early history would therefore have aided the preservation of the mantle's primordial C content. In addition, periods of upwelling of more-oxidized mantle, due to recycling or upwelling from a mainly isolated but more-oxidized lower mantle[8], may have extracted large amounts of CO_2 from the mantle in discrete episodes, such as those producing large igneous provinces[33].

METHODS SUMMARY

Experiments were performed using several starting materials to cover suitable ranges of f_{O_2} and to demonstrate the approach to equilibrium of garnet compositions from both initially reduced and oxidized (a reversal in terms of garnet oxidation state) garnet starting materials. Multi-anvil experimental capsules contained layers of natural garnet or synthetic glass representative of fertile garnet compositions (Supplementary Table 1) sandwiched between layers of a mineral mixture comprising orthopyroxene, garnet and olivine. In almost all the experiments, starting materials were placed inside graphite sleeves within a rhenium outer capsule. A flux of either carbonate or H_2O was added to all experiments to improve equilibration (Supplementary Table 2). The coexistence of graphite and carbonate melt in many experiments also buffered the oxygen fugacity[22] (Supplementary Fig. 1). One carbon-free experiment was performed in a Re metal capsule. Iridium metal powder (3 wt%) was added to all experiments, and alloyed with Fe to act as a redox sensor. Chemical compositions of the mineral phases were obtained using an electron microprobe. The equilibrium f_{O_2} in each experiment was calculated from the redox sensor using the expression

$$\log(f_{O_2}) = -\frac{\Delta G^0_{(2)}}{\ln(10)RT} + \log(a^{Ol}_{Fe_2SiO_4}) + \log(a^{Ol}_{Mg_2SiO_4}) - 2\log(a^{Opx}_{MgSiO_3}) - 2\log(a^{alloy}_{Fe}) \quad (5)$$

where T is the temperature, R is the gas constant, $\Delta G^0_{(2)}$ is the standard-state Gibbs free-energy change of equilibrium (2), a^{alloy}_{Fe} is the activity of Fe in the Ir-Fe alloy and $a^{Ol}_{Fe_2SiO_4}$ is the activity of the Fe_2SiO_4 component in olivine (and the other activities are defined similarly). The resulting $Fe^{3+}/\sum Fe$ ratio in garnets from high-pressure experiments was measured using Mössbauer spectroscopy. Samples were double-polished to a thickness of 200–300 μm. Spectra were collected at 298 K using a nominal 370-MBq ^{57}Co point source (Supplementary Table 3 and Supplementary Fig. 2).

The oxygen fugacity for equilibrium (3) was calculated from

$$\log(f_{O_2}) = -\frac{\Delta G^0_{(3)}}{\ln(10)RT} + 2\log(a^{Gt}_{Ca_3Fe_2Si_3O_{12}}) + 2\log(a^{Gt}_{Mg_3Al_2Si_3O_{12}}) + 4\log(a^{Opx}_{FeSiO_3}) - 2\log(a^{Gt}_{Ca_3Al_2Si_3O_{12}}) - 4\log(a^{Ol}_{Fe_2SiO_4}) - 6\log(a^{Opx}_{MgSiO_3}) \quad (6)$$

where $\Delta G^0_{(3)}$ is the standard-state Gibbs free-energy change of equilibrium (3). Activity–composition relations, thermodynamic data for equilibria (1)–(3) and calculated oxygen fugacities are reported in Supplementary Tables 4 and 5.

Full Methods and any associated references are available in the online version of the paper.

Received 22 June; accepted 15 October 2012.

1. Kasting, J. F., Eggler, D. H. & Raeburn, S. P. Mantle redox evolution and the oxidation state of the Archean atmosphere. *J. Geol.* **101,** 245–257 (1993).
2. Luth, R. W., Virgo, D., Boyd, F. R. & Wood, B. J. Ferric iron in mantle-derived garnets. *Contrib. Mineral. Petrol.* **104,** 56–72 (1990).
3. Woodland, A. B. & Koch, M. Variation in oxygen fugacity with depth in the upper mantle beneath Kaapvaal craton, South Africa. *Earth Planet. Sci. Lett.* **214,** 295–310 (2003).
4. O'Neill, H. St C. & Wall, V. J. The olivine-orthopyroxene-spinel oxygen geobarometer, the nickel precipitation curve, and the oxygen fugacity of the Earth's upper mantle. *J. Petrol.* **28,** 1169–1191 (1987).
5. Wood, B. J. Oxygen barometry of spinel peridotites. *Rev. Mineral. Geochem.* **25,** 417–432 (1991).
6. Ballhaus, C. & Frost, B. R. The generation of oxidized CO_2-bearing basaltic melts from reduced CH_4-bearing upper mantle sources. *Geochem. Cosmochem. Acta* **58,** 4931–4940 (1994).
7. Gudmundsson, G. & Wood, B. J. Experimental tests of garnet peridotite oxygen barometry. *Contrib. Mineral. Petrol.* **119,** 56–67 (1995).
8. Frost, D. J. & McCammon, C. A. The redox state of the Earth's mantle. *Annu. Rev. Earth Planet. Sci.* **36,** 389–420 (2008).
9. Ballhaus, C. Is the upper mantle metal-saturated? *Earth Planet. Sci. Lett.* **132,** 75–86 (1995).
10. O'Neill, H. St C., Rubie, D. C., Canil, D., Geiger, C. A. & Ross, C. R. in *Evolution of the Earth and Planets* (eds Takahashi, E., Jeanloz, R. & Rubie, D. C.) 74–88 (Geophys. Monogr. 74, American Geophysical Union, 1993).
11. Woodland, A. B. & O'Neill, H. St C. Thermodynamic data for Fe-bearing phases obtained using noble metal alloys as redox sensors. *Geochim. Cosmochim. Acta* **61,** 4359–4366 (1997).
12. Woodland, A. B. & O'Neill, H. St C. Synthesis and stability of $Fe_3Fe_2^{3+}Si_3O_{12}$ garnet and phase relations with $Fe_3Al_2Si_3O_{12}$-$Fe_3Fe_2^{3+}Si_3O_{12}$ solutions. *Am. Mineral.* **78,** 1000–1013 (1993).
13. Holland, T. & Powell, R. An improved and extended internally consistent thermodynamic dataset for phases of petrological interest, involving a new equation of state for solids. *J. Metamorph. Geol.* **29,** 333–383 (2011).
14. Canil, D. & O'Neill, H. St C. Distribution of ferric iron in some upper-mantle assemblages. *J. Petrol.* **37,** 609–635 (1996).
15. McCammon, C. A. & Kopylova, M. G. A redox profile of the Slave mantle and oxygen fugacity control in the cratonic mantle. *Contrib. Mineral. Petrol.* **148,** 55–68 (2004).
16. Lazarov, M., Woodland, A. B. & Brey, G. P. Thermal state and redox conditions of the Kaapvaal mantle: a study of xenoliths from the Finsch mine, South Africa. *Lithos* **112S,** 913–923 (2009).
17. Creighton, S. *et al.* Oxidation of the Kaapvaal lithospheric mantle driven by metasomatism. *Contrib. Mineral. Petrol.* **157,** 491–504 (2009).
18. Creighton, S., Stachel, T., Eichenberg, D. & Luth, R. W. Oxidation state of the lithospheric mantle beneath Diavik diamond mine, central Slave craton, NWT, Canada. *Contrib. Mineral. Petrol.* **159,** 645–657 (2010).
19. Yaxley, G. M., Berry, A. J., Kamenetsky, V. S., Woodland, A. B. & Golovin, A. V. An oxygen fugacity profile through the Siberian Craton-Fe K-edge XANES determinations of $Fe^{3+}/\Sigma Fe$ in garnets in peridotite xenoliths from the Udachnaya East kimberlite. *Lithos* **140–141,** 142–151 (2012).
20. Eggler, D. H. & Baker, D. R. in *High-Pressure Research in Geophysics* (eds Akimoto, S. & Manghnani, M. H.) 237–250 (Springer, 1982).
21. Dasgupta, R. & Hirschmann, M. M. The deep carbon cycle and melting in Earth's interior. *Earth Planet. Sci. Lett.* **298,** 1–13 (2010).
22. Stagno, V. & Frost, D. J. Carbon speciation in the asthenosphere: experimental measurements of the redox conditions at which carbonate-bearing melts coexist with graphite or diamond in peridotite assemblages. *Earth Planet. Sci. Lett.* **300,** 72–84 (2010).
23. Peslier, A. H., Woodland, A. B., Bell, D. R. & Lazarov, M. Olivine water contents in the continental lithosphere and the longevity of cratons. *Nature* **467,** 78–81 (2010).
24. McDonough, W. F. & Sun, S.-s. The composition of the Earth. *Chem. Geol.* **120,** 223–253 (1995).
25. Bézos, A. & Humler, E. The $Fe^{3+}/\sum Fe$ ratios of MORB glasses and their implications for mantle melting. *Geochim. Cosmochim. Acta* **69,** 711–725 (2005).
26. Blundy, J. D., Brodholt, J. P. & Wood, B. J. Carbon-fluid equilibria and the oxidation state of the upper mantle. *Nature* **349,** 321–324 (1991).
27. Saal, A. E., Hauri, E., Langmuir, C. H. & Perfit, M. R. Vapour undersaturation in primitive mid-ocean-ridge basalts and the volatile content of Earth's upper mantle. *Nature* **419,** 451–455 (2002).
28. Gaillard, F., Malki, M., Iacono-Marziano, G., Pichavant, M. & Scaillet, B. Carbonatite melts and electrical conductivity in the asthenosphere. *Science* **322,** 1363–1365 (2008).
29. Gu, Y. J., Lerner-Lam, A. L., Dziewonski, A. M. & Ekstrom, G. Deep structure and seismic anisotropy beneath the East Pacific Rise. *Earth Planet. Sci. Lett.* **232,** 259–272 (2005).
30. Rohrbach, A. *et al.* Metal saturation in the upper mantle. *Nature* **449,** 456–458 (2007).
31. Rohrbach, A. & Schmidt, M. W. Redox freezing and melting in the Earth's deep mantle resulting from carbon-iron redox coupling. *Nature* **472,** 209–212 (2011).
32. McKenzie, D. The extraction of magma from the crust and mantle. *Earth Planet. Sci. Lett.* **74,** 81–91 (1985).

33. Campbell, I. H. Large igneous provinces and the plume hypothesis. *Elements* **1,** 265–269 (2005).
34. O'Neill, H. St C. & Wood, B. J. An experimental study of Fe-Mg partitioning between garnet and olivine and its calibration as a geothermometer. *Contrib. Mineral. Petrol.* **70,** 59–70 (1979).
35. Balta, J. B., Asimow, P. D. & Mosenfelder, J. L. Hydrous, low-carbon melting of garnet peridotite. *J. Petrol.* **52,** 2079–2105 (2011).

Supplementary Information is available in the online version of the paper.

Acknowledgements Financial support was provided to V.S. by the European Commission under the Marie Curie Action for Early Stage Training of Researchers within the 6th Framework Programme (contract number MEST-CT-2005-019700) and by the German Science Foundation (grant FR1555/5-1).

Author Contributions V.S. and D.J.F. wrote the paper. V.S. performed most of the experiments, analytical measurements and calculations. D.O.O. performed high-pressure experiments. C.A.M. collected and interpreted Mössbauer data. D.J.F. developed the thermodynamic model.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to D.J.F. (dan.frost@uni-bayreuth.de).

METHODS

Starting materials and capsules. Starting materials for multi-anvil experiments consisted of layers of either natural garnet or garnet formed from synthetic glass (Supplementary Table 1) sandwiched between layers of natural olivine and orthopyroxene from San Carlos mixed with the same garnet composition. Compositions of starting garnets are given in Supplementary Table 1. Glasses were made from oxides melted and quenched at 1,600 °C. Two glasses (a and b in Supplementary Table 1) were then reduced at 800 °C in a H_2-CO_2 furnace at an oxygen fugacity equivalent to 2 log units less than the FMQ oxygen buffer. One of these glasses was then crystallized into two garnet samples in successive experiments performed at 3 GPa and 1,200 °C in a piston cylinder press using Pt capsules with added H_2O. Two resulting garnet compositions were more oxidized than the starting glass and had $Fe^{3+}/\sum Fe$ ratios of 0.05 and 0.1, respectively (e and f in Supplementary Table 1). A further glass composition (d in Supplementary Table 1) was not reduced and therefore preserved a high Fe_2O_3 content. In some experiments, 5% of a sub-equal mixture of orthopyroxene, olivine and carbonate was mixed within the garnet layer. In subsequent experiments using more-oxidized garnet glass compositions, only graphite was added to the garnet layer (these differences are indicated in Supplementary Table 2). All experiments except one (V721) were performed inside graphite sleeves and had 10% graphite powder mixed within the sandwiching olivine, orthopyroxene and garnet layers. Iridium metal (3 wt%) was added to all starting compositions to act as a redox sensor. In addition, a flux was added to all experiments. This was $MgCO_3$, $(Mg,Ca)CO_3$ (formed by adding a 1:1 mixture of $MgCO_3$ and $CaSiO_3$, with the latter also resulting in the formation of diopside) or H_2O added as brucite. The graphite sleeves were enclosed in Re-foil outer capsules[22]. Experiment V721 was graphite free and was encapsulated in a Re sleeve inside a Pt capsule.

Multi-anvil experiments. Multi-anvil experiments were performed using tungsten carbide cubes with a corner-truncation edge length of 11 mm, in Cr_2O_3-doped MgO octahedral pressure media with an edge length of 18 mm. $LaCrO_3$ or graphite heaters were used, placed inside an outer ZrO_2 sleeve. Experimental capsules were 2 mm in diameter and 3.5 mm in length, and were placed at the centre of the cylindrical heater inside MgO spacers. The temperature was monitored with a $W_{97}Re_3$–$W_{75}Re_{25}$ (D-type) thermocouple inserted within an alumina sleeve, with the junction in contact with the top of the capsule. Further details can be found in ref. 22.

Analytical techniques. All the recovered run products were mounted in epoxy resin and then ground and polished using ethanol to preserve the carbonate phases (solid and melt) for analytical investigations. The chemical compositions of liquid and mineral phases were obtained using a Jeol JXA-8200 electron microprobe equipped with five wavelength-dispersive spectrometers. The $Fe^{3+}/\sum Fe$ ratio in garnets recovered from high-pressure experiments was measured using Mössbauer spectroscopy. Samples were double-polished to a thickness of approximately 200–300 μm to give an absorber thickness of about 5 mg Fe cm^{-2} and to avoid saturation effects. Spectra were collected at 298 K using a nominal 370-MBq ^{57}Co point source[36]. A velocity range of −5 to +5 mm s^{-1} was always used. The velocity scale was calibrated relative to a 25-μm-thick α-Fe foil. Finally, spectra were folded and fitted to Lorentzian line shapes using the fitting program MOSSA[37]. Spectra were fitted using one doublet for ferrous iron in the dodecahedral site, one doublet for ferric iron in the octahedral site[2,38] and one doublet (where needed) for Fe^{2+} in olivine and/or orthopyroxene, using conventional constraints. The ferric iron content in garnets was calculated from the relative areas corrected for differences in recoil-free fraction[38] and, together with hyperfine parameters, are reported in Supplementary Tables 2 and 3.

Thermodynamic data and activity models. The oxygen fugacity in each experiment was determined from equilibrium (2) using the Ir-Fe redox sensor[11,22,39,40] and equation (5). Thermodynamic data to calculate $\Delta G^0_{(2)}$ are reported in ref. 22. The oxygen fugacity determined for the garnet peridotite oxythermobarometer equilibrium (3) is calculated using equation (6). We determined $\Delta G^0_{(3)}$ from the data of ref. 13, and fitted it to

$$\frac{\Delta G^0_{(3)}}{\ln(10)RT} = -8.6623 + \frac{21{,}655}{T} + \frac{(13.6149 - 0.01833T)P}{T}$$

with temperature in kelvin and pressure in kilobars.

Activity–composition relations for components in both equations are described below, with all interaction terms reported in Supplementary Table 4.

The activities of Fe_2SiO_4 and Mg_2SiO_4 components in olivine are described by

$$a^{Ol}_{Mg_2SiO_4} = \left(X_{Mg}\gamma_{Mg}\right)^2$$

$$a^{Ol}_{Fe_2SiO_4} = \left(X_{Fe}\gamma_{Fe}\right)^2$$

where $X_{Fe} = Fe/(Fe+Mg)$ and the activity coefficients γ_m (m = Fe, Mg) are described using

$$RT\ln(\gamma_m) = W^{Ol}_{Fe\text{-}Mg}(1 - X_m)^2$$

where $W^{Ol}_{Fe\text{-}Mg}$ is a symmetric Margules interaction parameter[41].

The activities of $FeSiO_3$ and $MgSiO_3$ in orthopyroxene are assumed to be ideal, for example

$$a^{Opx}_{FeSiO_3} = \left(X_{Fe}\right)$$

and the activity of Fe in Ir-Fe alloy is described by the expression

$$a^{alloy}_{Fe} = \left(X_{Fe}\gamma_{Fe}\right)$$

where

$$RT\ln(\gamma_{Fe}) = (1 - X_{Fe})^2\left[W^{alloy}_{Ir\text{-}Fe} + 2\left(W^{alloy}_{Ir\text{-}Fe} - W^{alloy}_{Ir\text{-}Fe}\right)\left(X_{Fe}\right)\right]$$

For garnets, the activity of each component is described by

$$a^{Gt}_{Mg_3Al_2Si_3O_{12}} = \left(X_{Mg}\gamma_{Mg}\right)^3_{dodec}\left(X_{Al}\gamma_{Al}\right)^2_{oct}\left(\gamma_{Mg_3Al_2Si_3O_{12}}\right)_{reciprocal}$$

$$a^{Gt}_{Ca_3Al_2Si_3O_{12}} = \left(X_{Ca}\gamma_{Ca}\right)^3_{dodec}\left(X_{Al}\gamma_{Al}\right)^2_{oct}\left(\gamma_{Ca_3Al_2Si_3O_{12}}\right)_{reciprocal}$$

$$a^{Gt}_{Ca_3Fe^{3+}_2Si_3O_{12}} = \left(X_{Ca}\gamma_{Ca}\right)^3_{dodec}\left(X_{Fe^{3+}}\gamma_{Fe^{3+}}\right)^2_{oct}\left(\gamma_{Ca_3Fe^{3+}_2Si_3O_{12}}\right)_{reciprocal}$$

The mole fraction of Mg on the dodecahedral site of garnet is, for example Mg/(Mg+Ca+Fe^{2+}), and the mole fraction of Al on the octahedral site of garnet is Al/(Al+Fe^{3+}+Cr). Activity coefficients are calculated using ternary Margules activity–composition expressions[42], with symmetric[2] and asymmetric[42] terms used for octahedral and dodecahedral sites, respectively. Non-ideality due to multisite mixing in garnet, that is, $(\gamma_{Mg_3Al_2Si_3O_{12}})_{reciprocal}$, is treated with respect to the following reciprocal reactions[43]:

$$Mg_3Al_2Si_3O_{12} + Ca_3Fe_2Si_3O_{12} = Mg_3Fe_2Si_3O_{12} + Ca_3Al_2Si_3O_{12} \quad (7)$$

$$Fe_3Al_2Si_3O_{12} + Ca_3Fe_2Si_3O_{12} = Fe_3Fe_2Si_3O_{12} + Ca_3Al_2Si_3O_{12} \quad (8)$$

$$Mg_3Al_2Si_3O_{12} + Ca_3Cr_2Si_3O_{12} = Mg_3Cr_2Si_3O_{12} + Ca_3Al_2Si_3O_{12} \quad (9)$$

$$Fe_3Al_2Si_3O_{12} + Ca_3Cr_2Si_3O_{12} = Fe_3Cr_2Si_3O_{12} + Ca_3Al_2Si_3O_{12} \quad (10)$$

The standard-state Gibbs free-energy changes of these reactions (as indicated by subscripts) are determined to be $\Delta G^0_{(7)} = \Delta G^0_{(8)} = 33{,}000 - 70P$ J (where P is expressed in kilobars; ref. 12) and $\Delta G^0_{(9)} = \Delta G^0_{(10)} = 50{,}000$ J (ref. 7). The corresponding activity coefficients arising from these reciprocal expressions are[43]

$$RT\ln\left(\gamma_{Ca_3Al_2Si_3O_{12}}\right)_{reciprocal} = -X_{Mg}X_{Fe^{3+}}\Delta G^0_{(7)} - X_{Fe}X_{Fe^{3+}}\Delta G^0_{(8)} - X_{Mg}X_{Cr}\Delta G^0_{(9)} - X_{Fe}X_{Cr}\Delta G^0_{(10)}$$

$$RT\ln\left(\gamma_{Ca_3Fe_2Si_3O_{12}}\right)_{reciprocal} = X_{Mg}(1 - X_{Fe^{3+}})\Delta G^0_{(7)} + X_{Fe}(1 - X_{Fe^{3+}})\Delta G^0_{(8)} - X_{Mg}X_{Cr}\Delta G^0_{(9)} - X_{Fe}X_{Cr}\Delta G^0_{(10)}$$

$$RT\ln\left(\gamma_{Mg_3Al_2Si_3O_{12}}\right)_{reciprocal} = (1 - X_{Mg})X_{Fe^{3+}}\Delta G^0_{(7)} - X_{Fe}X_{Fe^{3+}}\Delta G^0_{(8)} + (1 - X_{Mg})X_{Cr}\Delta G^0_{(9)} - X_{Fe}X_{Cr}\Delta G^0_{(10)}$$

Uncertainties for the alloy sensor are ~0.3 log units, taking into account uncertainties in pressure and temperature from experimental runs (0.5 GPa and 50 °C), from the chemical composition of the run products and from thermodynamic data[22].

The contribution of the reciprocal reactions to the activity coefficient terms is the main source of uncertainty in the use of equilibrium (3) to determine f_{O_2}. In particular, there are large uncertainties in the magnitudes $\Delta G^0_{(7)}$ and $\Delta G^0_{(8)}$, which here are assumed to be equal and constrained by a previous study[12]. Taking into account the errors in the previous determinations of these values, we arrive at an uncertainty of 10 kJ mol^{-1} for $\Delta G^0_{(7)}$ and $\Delta G^0_{(8)}$, which propagates to an error of approximately 0.5 log units in f_{O_2}. Combining this error with that typical for garnet ferric Fe determinations results in an uncertainty of 0.7 log units.

Origin of the discrepancy between f_{O_2} determinations using equilibria (1) and (3). As shown in Fig. 1, oxygen fugacities determined using equilibrium (1) are lower than those calculated with equilibrium (3) and the Ir-Fe redox sensor at pressures of 6–7 GPa. As stated, we note that thermodynamic data for the $Fe_3Fe_2^{3+}Si_3O_{12}$ garnet component skiagite are sparse and are determined from a previous study that examined the equilibrium

$$\underset{\text{garnet}}{Fe_3Fe_2^{3+}Si_3O_{12}} + \underset{\text{spinel}}{FeAl_2O_4} = \underset{\text{garnet}}{Fe_3Al_2Si_3O_{12}} + \underset{\text{spinel}}{Fe_3O_4}$$

and determined ΔG^0 for this exchange equilibrium at 1,300 K (ref. 12). Using ancillary thermodynamic data, it was then possible to estimate $\Delta_f G^0$, the Gibbs free energy of formation, of the $Fe_3Fe_2^{3+}Si_3O_{12}$ garnet component at 1,373 K. However, no data exist with which to determine the entropy for this component, which was therefore estimated in ref. 7. As a result, large uncertainties must be associated with the Gibbs free-energy change for equilibrium (1) at temperatures other than 1,373 K. Uncertainties in this estimate might at least partly explain the difference in calculated oxygen fugacities for equilibria (1) and (3) at high pressures. Although the reciprocal reactions for equilibrium (3) also require thermodynamic data for $Fe_3Fe_2^{3+}Si_3O_{12}$ and $Mg_3Fe_2^{3+}Si_3O_{12}$, which are assumed to have identical values, here uncertainties in the entropy are less important as they probably cancel out on either side of the equilibrium, enabling the estimated $\Delta_f G^0$ for $Fe_3Fe_2^{3+}Si_3O_{12}$ to provide a sufficient approximation for the Gibbs free-energy changes of the reciprocal reactions.

Calculation of the oxygen fugacity along a mantle adiabat. The calculation of mantle f_{O_2} for a bulk silicate Earth composition as a function of bulk rock $Fe^{3+}/\sum Fe$ ratio (Fig. 3) is made using equation (6). Mineral compositions for a four-phase assemblage are determined by performing a mass balance with Al contents of orthopyroxene and clinopyroxene determined from experimental and natural data[2,14,35] and using experimental Fe-Mg partitioning data[34,35]. The distribution of Fe^{3+} between garnet, clinopyroxene and orthopyroxene is determined from models derived from analyses of natural xenoliths[14]. No consideration of the majorite component in garnet was made because the proportion should be small at the depths considered in the calculation. Similarly the orthopyroxene to clinoenstatite (the high-pressure monoclinic polymorph with space group $C2/c$) transition was ignored.

36. McCammon, C. A. A Mössbauer milliprobe: practical considerations. *Hyperfine Interact.* **92,** 1235–1239 (1994).
37. Prescher, C., McCammon, C. & Dubrovinsky, L. MossA: a program for analyzing energy-domain Mossbauer spectra from conventional and synchrotron sources. *J. Appl. Crystallogr.* **45,** 329–331 (2012).
38. Amthauer, G., Annersten, H. & Hafner, S. S. The Mössbauer spectrum of ^{57}Fe in silicate garnets. *Z. Kristallogr.* **143,** 14–55 (1976).
39. Schwerdtfeger, K. & Zwell, L. Activities in solid iridium-iron and rhodium-iron alloys at 1200°C. *Trans. Metall. Soc. AIME* **242,** 631–633 (1968).
40. Swartzendruber, L. J. The Fe-Ir (iron-iridium) system. *Bull. Alloy Phase Diagr.* **5,** 48–52 (1984).
41. Wiser, N. & Wood, B. J. Experimental determination of activities in Fe-Mg olivine at 1400K. *Contrib. Mineral. Petrol.* **108,** 146–153 (1991).
42. Ganguly, J., Cheng, C. & Tirone, M. Thermodynamics of aluminosilicate garnet solid solutions: new experimental data, an optimized model, and thermometric applications. *Contrib. Mineral. Petrol.* **126,** 137–151 (1996).
43. Wood, B. J. & Nicholls, J. The thermodynamical properties of reciprocal solid solutions. *Contrib. Mineral. Petrol.* **66,** 389–400 (1978).

LETTER

doi:10.1038/nature11777

Ediacaran life on land

Gregory J. Retallack[1]

Ediacaran (635–542 million years ago) fossils have been regarded as early animal ancestors of the Cambrian evolutionary explosion of marine invertebrate phyla[1], as giant marine protists[2] and as lichenized fungi[3]. Recent documentation of palaeosols in the Ediacara Member of the Rawnsley Quartzite of South Australia[4] confirms past interpretations of lagoonal–aeolian deposition based on synsedimentary ferruginization and loessic texture[5,6]. Further evidence for palaeosols comes from non-marine facies, dilation cracks, soil nodules, sand crystals, stable isotopic data and mass balance geochemistry[4]. Here I show that the uppermost surfaces of the palaeosols have a variety of fossils in growth position, including *Charniodiscus*, *Dickinsonia*, *Hallidaya*, *Parvancorina*, *Phyllozoon*, *Praecambridium*, *Rugoconites*, *Tribrachidium* and 'old-elephant skin' (ichnogenus *Rivularites*[7]). These fossils were preserved as ferruginous impressions, like plant fossils[8], and biological soil crusts[9,10] of Phanerozoic eon sandy palaeosols. Sand crystals after gypsum[11] and nodules of carbonate[12] are shallow within the palaeosols[4], even after correcting for burial compaction[13]. Periglacial involutions and modest geochemical differentiation of the palaeosols are evidence of a dry, cold temperate Ediacaran palaeoclimate in South Australia[4]. This new interpretation of some Ediacaran fossils as large sessile organisms of cool, dry soils, is compatible with observations that Ediacaran fossils were similar in appearance and preservation to lichens and other microbial colonies of biological soil crusts[3], rather than marine animals[1], or protists[2].

Newly documented palaeosols in the Ediacara Member of the Rawnsley Quartzite in South Australia[4] now call for a re-evaluation of its famous fossils, widely considered evolutionary predecessors of the Cambrian explosion of marine animal phyla[1]. Ediacaran red beds of South Australia (Figs 1 and 2) were initially considered non-marine by Douglas Mawson and Ralph Segnit[5]. When Mawson's student Reginald Sprigg discovered and interpreted South Australian Ediacaran fossils as marine jellyfish[14], this palaeoenvironmental contradiction was resolved by a compromise interpretation of jellyfish thrown up onto tidal flats by storms[15]. Ediacaran fossils are known worldwide in a variety of sedimentary facies[16], generally interpreted as shallow to deep marine, following Sprigg's[14] comparison with marine animals[1], although such comparisons now seem increasingly doubtful. Most Ediacaran fossils have no clear relationship with modern animals[2,3,16,17]. Putative Neoproterozoic 'embryos' were more likely to have been protists[18]. Putative permineralized metazoans may instead have been crystal-lined vughs[19], and other permineralized Ediacaran fossils were red algae or glomeromycotan lichens[20]. Precambrian shallow trails may have been made by slime moulds in their slug aggregation phase rather than worms[21]. There have also been suggestions that Ediacaran fossils were giant protists, such as xenophyopores[2], or fungi, such as lichens[3].

Palaeosols in the Ediacara Member have been overlooked until now, because they are less strongly developed than palaeosols at Precambrian unconformities or formed under forests of the Devonian period and later[4]. In addition to obvious soil structures (platy peds) and horizons (A-Bk and A-By), bedding disruption on mainly microscopic scales contributes to the massive appearance of palaeosols compared with sedimentary rocks in the field (Fig. 2b–e). One bed (Warrutu palaeosol of Fig. 2b) has four distinct episodes of soft sediment deformation followed by successive weathering and bedding disruption of previous episodes, comparable with successive periglacial soil involutions, rather than seismic or load deformation[4]. Further evidence for Ediacaran palaeosols, detailed elsewhere[4] includes (1) geochemical mass-balance negative strain and cation mass transfer; (2) loessic grain-size distribution and texture; (3) unusually light carbon and oxygen isotopic compositions that show linear covariance; and (4) sand crystals of gypsum and micritic replacive nodules with a consistent depth from the tops of beds. The 47 different stratigraphic levels showing pedogenic features in Brachina Gorge (Fig. 1) are repetitions of five distinct types of palaeosol named as pedotypes from the Adnamatna indigenous language (see Supplementary Information, Supplementary Fig. 2 and Supplementary Tables 1 and 2).

The red colour and weathering of rocks in the Flinders Ranges have been regarded as products of deep weathering from the Cretaceous period or later[22], but this view is falsified by several observations. Red beds of the Ediacara Member have been found beneath grey shales and limestones in drill cores in the Ediacara hills, where the cores have the same unusual and distinctive carbonate carbon isotopic composition and major element composition as the outcrops in Brachina Gorge[4]. Grey sandstone palaeochannels of the Ediacara Member include red clasts redeposited from the Ediacara Member and Bonney Sandstone, as well as grey calcareous clasts from the Wonoka Formation: all should be red if they were weathered downwards from the current land surface. A variety of clay crystallinity indices, X-ray diffraction data and microprobe analyses demonstrate that the Ediacara Member in outcrop and core is illite-chlorite that has been heated to low within the greenschist metamorphic facies[4] (see Supplementary Information, Supplementary Figs 3–6 and Supplementary Table 4). Furthermore, Ediacara Member sandstones contain abundant feldspar, and red siltstones include carbonate nodules. By contrast, deep weathering profiles contain no feldspar nor carbonate, and have kaolinite clays and a chemical composition strongly depleted in alkali and alkaline earth elements[4]. The red colour and degree of weathering do not distinguish palaeosols from sediments, because both can be acquired from soils in source terrains, but the relative timing of reddening and weathering is crucial to a terrestrial–aeolian interpretation[5].

Palaeosols and fossils of the Ediacara Member are distinguished by a surface texture called old-elephant skin, which is best preserved under covering sandstone beds[7]. *Rivularites repertus* is a validly named ichnospecies for comparable cracked and pustulose surfaces[7]. What makes it look old is healed cracks, irregular fine ridges (cracks in cover sandstone), and pustulose relief of intergrown radial growth centres (Fig. 2a). These distinctive cracked and pustulose surfaces have a variety of features that are more like the biological soil crusts of desert and tundra[9,10] than the parallel-wrinkled, and undulose hydrated microbial mats of intertidal flats and shallow seas[7]. Biological soil crusts and their soils have vertically oriented organisms intimately admixed with minerals of the soil, whereas aquatic microbial mats are laminated, and detachable from their mineral substrate as flakes, skeins and rollups, not seen in the Ediacara Member. Soil crusts have irregular relief, healed desiccation cracks and pressure ridges even in clay-poor sandstones, whereas microbial mats have flexuous, striated domes and tufts, again not seen in the Ediacara Member. Soil crusts are

[1]Department of Geological Sciences, University of Oregon, Eugene, Oregon 97403-1272, USA.

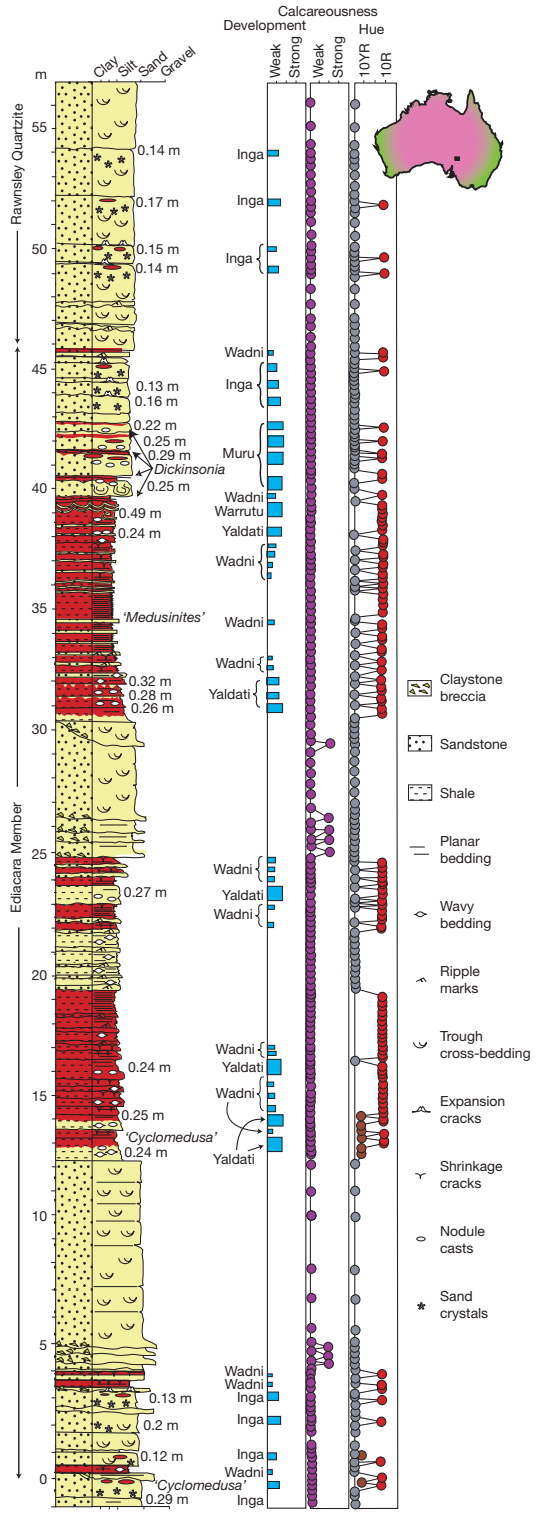


Figure 1 | Geological section of upper Ediacara Member in Brachina Gorge, South Australia. The interpreted palaeosol position and development (height and width of black boxes, respectively[4]) are shown. Calcareousness assessed in field by degree of reaction with 10% stock HCl. Hues such as 10YR and 10R are from a Munsell chart. This was the entire Ediacara Member as originally defined[6] in the Brachina Gorge (31.34422° S, 138.55763° E).

the upper part of deeper soil profiles with downward variation in oxidation, clay abundance and replacive nodular subsurface horizons like palaeosols of the Ediacara Member, whereas microbial mats form caps to unweathered, chemically reduced sedimentary layers. Soil crusts develop increasingly differentiated soil profiles through time, whereas microbial mats build upwards in laminar to domed (stromatolitic) increments. Sandy palaeosols with impressions of lichen-like fossils are also known from Ordovician[9] and Cretaceous palaeosols[10]. Comparable preservation of vascular land plants is well known in red sandy palaeosols of Cretaceous age[8].

Ediacaran fossils were preserved as impressions in old-elephant skin sandstones overlying four of the five different kinds of palaeosol at ten different horizons in four classic localities (see Supplementary Information and Supplementary Table 3). The best place to see Ediacaran fossils in place above a palaeosol is the overhang at the 39.7 m level in Brachina Gorge (Fig. 2b), where there are still good specimens of *Dickinsonia costata* and *Pseudorhizostomites howchini*. The body fossils were firmly attached or embedded within these soil surfaces in life, as revealed by growth series, lack of overlapping specimens, and thickening of adjacent specimens comparable with competitive reaction[3,22]. Some taxa such as *Phyllozoon hanseni* and *Aulozoon* sp. are embedded within the surface layer, like window lichens and rhizines in desert crusts[3]. In petrographic thin sections, branching tubular structures extend deep into the palaeosols (Fig. 2c–g), like cyanobacterial ropes, fungal hyphae and lichen rhizines of desert crust soils today[7]. Petrographic thin sections of the lower half of impressions of *Dickinsonia* fossils show comparable bedding disruption by irregular tubular features (Fig. 2c, d): the more common upper impression fossils show only bedded sandstone overlying the fossil (Fig. 2e). Comparable bifacial fossil features, with smooth and finished upper surfaces but ragged lower surfaces, were also found in surface horizons of the palaeosols (Fig. 2e–g). Which specific Ediacaran fossils are represented by these thin sections is uncertain, because Ediacaran taxa have been defined by shape, not appearance in thin section. These images of complex chambered structures with basal tubules (Fig. 2c–g) are preliminary indications of their appearance in thin section prepared for a detailed study in progress. Observed cross-sections of Ediacaran fossils in petrographic thin sections are comparable in preservational style with plants and lichens in Phanerozoic palaeosols[8–10].

Body fossils so far documented on the palaeosols include '*Aulozoon*' sp. indet., *Charniodiscus arboreus*, '*Cyclomedusa davidi*', *Dickinsonia costata*, *D. elongata*, *D. rex.*, *Hallidaya brueri*, cf. '*Kimberella*' sp. indet., '*Medusinites asteroides*', *Parvancorina minchami*, *Phyllozoon hanseni*, *Praecambridium sigillum*, *Rugoconites enigmaticus*, *Spriggina floundersi* and *Tribrachidium heraldicum*. Trace fossils found in the palaeosols include *Archaeonassa* sp. indet., *Pseudorhizostomites howchini*, cf. '*Radulichnus*' sp. indet. and *Rivularites repertus* (see Supplementary Information and Supplementary Table 3). Some palaeosols (Muru and Wadni pedotypes) have a diverse fossil assemblage dominated by *Dickinsonia*, whereas others (Yaldati and Inga) have a low-diversity assemblage mainly of discoid fossils (variously attributed to microbial colonies, medusoids or holdfasts, and taxa such as *Medusinites* and *Cyclomedusa*[1]). One palaeosol with deformation interpreted as periglacial involutions[4] (Warrutu pedotype) yielded no fossil specimens in outcrops, but may have Ediacaran fossils in thin sections (Fig. 2f, g). These differences in diversity may be evidence that Ediacaran organisms preferred unfrozen, low salinity soils, rich in nutrients, like most terrestrial organisms.

Not only fossil diversity, but fossil size may have increased with palaeosol development. The relationship between gypsum abundance (G in area (%)) and geological age (A in kyr) in the Sinai and Negev Deserts of Israel[23,24] is given by the following equation:

$$A = 3.987G + 5.774 \quad (1)$$

This relationship ($R^2 = 0.95$) has the standard error ± 15 kyr. The largest *Dickinsonia* (32 cm long, *D. rex*[12]) from the main collecting gully in the Ediacara hills came from a Muru palaeosol with 10% gypsum, and using equation (1), was 53.6 ± 15 kyr in the making. Long-term growth of this *D. rex* (0.006 ± 0.002 mm yr^{-1}) would have been more like modern lichens (*Rhizocarpon geographicum*) in the Brooks Range

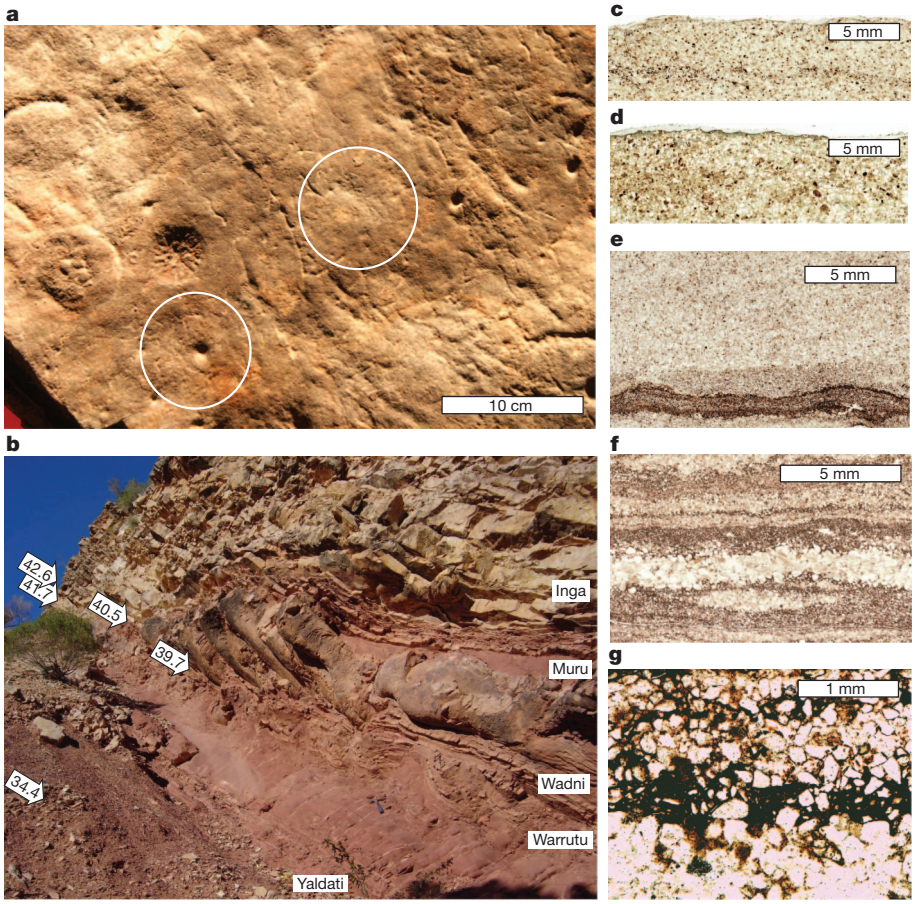


Figure 2 | Palaeosols of the Ediacara Member of the Rawnsley Quartzite, South Australia. **a**, *Rivularites repertus* (old-elephant skin) showing sutured radial growth, crack fills and ridge impressions, effaced discoid fossils (white circles) and fossil impressions (*Hallidaya brueri* in positive relief to left, and *Rugoconites enigmaticus* in negative relief to right), on sole of sandstone slab from Crisp Gorge (31.176572° S, 138.328533° E). **b**, Palaeosols and fossil levels in the Brachina Gorge (31.34422° S, 138.55763° E). **c**, **d**, Vertical petrographic thin sections of the lower part of *D. elongata* (**c**) and *D. costata* (**d**) showing bedding disruption and tubular features, from Muru palaeosol in the Ediacara hills (**c**) and unknown palaeosol at the Hookapunna well (**d**). **e**–**g**, Unidentified dorsoventral Ediacaran fossils in thin sections showing overlying cross-bedded sandstone (**e**) and basal irregular tubular structures (**e**–**g**). The fossiliferous surface in **a** is part of a large slab on display in the South Australian Museum, Adelaide. The hammer for scale in **b** has a length of 25 cm. Specimen numbers in the Condon Collection, Museum of Natural and Cultural History, University of Oregon are F112999 from Muru palaeosol in the Ediacara hills (**c**), F115736 from unknown palaeosol near the Hookapunna well (**d**), R3218 from Muru palaeosol in the Brachina Gorge (**e**), and both R3223 (**f**) and R3222 (**g**) are from the Warrutu palaeosol in the Brachina Gorge. All thin sections were cut vertical to regional bedding.

of northern Alaska (0.04 mm yr^{-1}) than lichens in southern Alaska[25] (0.1 mm yr^{-1}) or southern Norway[26] (0.5–0.7 mm yr^{-1}). Correlation between the largest *Dickinsonia* found within a collection and gypsum enrichment in the palaeosol of that collection is evidence for slow non-linear growth rates of *Dickinsonia* (Fig. 3, Supplementary Information and Supplementary Table 3). The Ediacara hills has yielded one of the most diverse assemblages known, and Ediacaran organisms may have diversified as well as grown with soil age, comparable with modern terrestrial communities.

Indications of palaeoclimate come from comparison of palaeosols in the Ediacara Member with modern soils. The closest modern analogue to Ediacaran gypsic palaeosols seems to be soilscapes on the coastal plain of the Caspian Sea near Atyrau, Kazakhstan (mean annual temperature 8 °C, mean annual precipitation 160 mm: map unit Zo16-3a of Orthic Solonchak, with associated Takyric and Gleyic Solonchaks[27]). Calcic palaeosols are found nearby in the Emba River floodplain (map units Xl 16-1ab and Jc 53-2c). Periglacial involutions in one palaeosol[4] support other evidence from dropstones for cold temperate palaeoclimate in the coeval Billy Springs Formation of the far northern Flinders Ranges[28]. Other indications of cool and dry palaeoclimate from climofunctions for modern soils are outlined in the Supplementary Information, Supplementary Fig. 7 and Supplementary Tables 5 and 6. The zone of best preservation of large Ediacaran fossils in the Flinders Ranges is thus near the gypsic–calcic soil ecotone, an important biotic boundary in modern temperate deserts, such as the Atacama[29] and Negev[23] Deserts.

Discovery of some Ediacaran fossils in the surface horizons of palaeosols does not mean that all Ediacaran fossils everywhere were

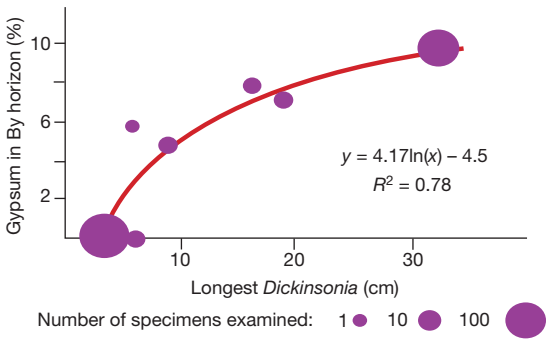


Figure 3 | Maximum length of *Dickinsonia* fossils related to area (%) of gypsum in same palaeosol, as a proxy for soil development. Measurements of the longest specimen from the largest collections are the most reliable (see Supplementary Information for data).

terrestrial. Intertidal to shallow marine facies recognized in the Ediacara Member[4], have so far proven unfossiliferous, but *Palaeopascichnus* is known from both old-elephant skin surfaces of the Ediacara Member interpreted here as terrestrial, as well as in pyritic black shales of the Wonoka Formation interpreted as marine[22]. Multisegmented fossils from palaeosols of the Ediacara Member such as *Dickinsonia*, *Charnia*, *Praecambridium* and *Spriggina* are more likely to have been lichens or other microbial consortia[3] than marine invertebrates[1] or giant protists[2]. Discoid Ediacaran fossils such as *Cyclomedusa*, *Medusinites* and *Rugosoconites* would not be jellyfish in such dry soils, but could have been microbial colonies[30]. Small fossils such as *Parvancorina* or *Tribrachidium* could not have been pre-trilobites or proto-seat-stars[1], respectively, if they lived on land, but may have been fungal-fruiting bodies[3]. Trace fossils such as *Archaeonassa* could have been created by metazoan slugs or worms after rainstorms on land, but terrestrial habitats also open the possibility that these trails were created by slug-aggregating phases of slime moulds[21]. '*Radulichnus*' impressions from the Ediacara Member are too straight and sharp to be molluscan radular scratches[2], and in cool temperate soils may instead have been casts of needle ice. *Pseudorhizostomites* has been considered a gas-escape structure in a marine setting[16], but as a soil-surface feature it is most like a flanged pedestal of a biological soil crust[7]. These surprising alternative terrestrial hypotheses for habitats and affinities of these enigmatic fossils arise largely from recognition of palaeosols, and their interpretation by comparison with modern soils and soil processes. These unconventional ideas and comparisons remain to be tested for different kinds of Ediacaran fossil, and in sequences and assemblages of Ediacaran fossils beyond South Australia.

METHODS SUMMARY

The main contribution of this Letter is to document the geographic and stratigraphic occurrence of Ediacaran fossils in palaeosols described in detail elsewhere[4], as well as new observations of Ediacaran fossils in petrographic thin sections. Further analytical data on these palaeosols are provided in the Supplementary Information and Supplementary Tables 4–6, and interpretation of their palaeoenvironmental setting is in Supplementary Tables 1, 2 and 7–9 and Supplementary Figs 2 and 7.

Received 11 July; accepted 9 November 2012.
Published online 12 December 2012.

1. Erwin, D. H. *et al.* The Cambrian conundrum: early divergence and later ecological success in the history of animals. *Science* **334,** 1091–1097 (2011).
2. Seilacher, A., Buatois, L. A. & Mangano, M. G. Trace fossils in the Ediacaran–Cambrian transition: behavioral diversification, ecological turnover and environmental shift. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **227,** 323–356 (2005).
3. Retallack, G. J. Growth, decay and burial compaction of *Dickinsonia*, an iconic Ediacaran fossil. *Alcheringa* **31,** 215–240 (2007).
4. Retallack, G. J. Were Ediacaran siliciclastics of South Australia coastal or deep marine? *Sedimentology* **59,** 1208–1236 (2012).
5. Mawson, D. & Segnit, E. R. Purple slates of the Adelaide System. *Trans Roy. Soc. S. Australia* **72,** 276–280 (1949).
6. Jenkins, R. J. F., Ford, C. H. & Gehling, J. G. The Ediacara Member of the Rawnsley Quartzite: the context of the Ediacara assemblage (late Precambrian, Flinders Ranges). *J. Geol. Soc. Australia* **30,** 101–119 (1983).
7. Retallack, G. J. Criteria for distinguishing microbial mats and earths. *Soc. Econ. Paleont. Mineral. Spec. Pap.* **101,** 136–152 (2012).
8. Retallack, G. J. & Dilcher, D. L. Core and geophysical logs versus outcrop for interpretation of Cretaceous paleosols in the Dakota Formation of Kansas. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **329–330,** 47–63 (2012).
9. Retallack, G. J. Cambrian–Ordovician non-marine fossils from South Australia. *Alcheringa* **33,** 355–391 (2009).
10. Simpson, W. S. *et al.* A preserved Late Cretaceous biological soil crust in the capping sandstone member, Wahweap Formation, Grand Staircase-Escalante National Monument, Utah: paleoclimatic implications. *Sedim. Geol.* **230,** 139–145 (2010).
11. Retallack, G. J. & Huang, C.-M. Depth to gypsic horizon as a proxy for paleoprecipitation in paleosols of sedimentary environments. *Geology* **38,** 403–406 (2010).
12. Retallack, G. J. Pedogenic carbonate proxies for amount and seasonality of precipitation in paleosols. *Geology* **33,** 333–336 (2005).
13. Sheldon, N. D. & Retallack, G. J. Equation for compaction of paleosols due to burial. *Geology* **29,** 247–250 (2001).
14. Sprigg, R. C. Early Cambrian (?) jellyfishes from the Flinders Ranges, South Australia. *Trans Roy. Soc. S. Australia* **71,** 212–224 (1947).
15. Glaessner, M. F. Precambrian animals. *Sci. Am.* **204,** 72–78 (1961).
16. Fedonkin, M. A., Gehling, J. G., Grey, K., Narbonne, G. M. & Vickers-Rich, P. *The Rise of Animals: Evolution and Diversification of the Kingdom Animalia* (Johns Hopkins Univ. Press, 2008).
17. Antcliffe, J. B. & Brasier, M. D. *Charnia* at 50: developmental models for Ediacaran fronds. *Palaeontology* **51,** 11–26 (2008).
18. Huldtgren, T. *et al.* Fossilized nuclei and germination structures identify Ediacaran "animal embryos" as encysting protists. *Science* **334,** 1696–1699 (2011).
19. Yin, Z. *et al.* Early embryogenesis of potential bilaterian animals with polar lobe formation from the Ediacaran Weng'an Biota, South China. *Precambr. Res.* http://dx.doi.org/10.1016/j.precamres.2011.08.011 (9 September 2011).
20. Yuan, X.-L., Xiao, S.-H. & Taylor, T. N. Lichen-like symbiosis 600 million years ago. *Science* **308,** 1017–1020 (2005).
21. Bengtson, S., Rasmussen, B. & Krapež, B. The Paleoproterozoic megascopic Stirling biota. *Paleobiology* **33,** 351–381 (2007).
22. Gehling, J. G., Droser, M. L., Jensen, S. R. & Runnegar, B. N. in *Evolving Form and Function: Fossils and Development* (ed. Briggs, D. E. G.) 45–56 (Yale Peabody Museum, 2005).
23. Dan, J., Moshe, R. & Alperovich, N. The soils of Sede Zin. *Israel J. Earth Sci.* **22,** 211–227 (1973).
24. Dan, J., Yaalon, D. H., Moshe, R. & Nissim, S. Evolution of reg soils in southern Israel and Sinai. *Geoderma* **28,** 173–202 (1982).
25. Solomina, O. & Calkin, P. E. Lichenometry as applied to moraines in Alaska, USA, and Kamchatka, Russia. *Arct. Antarct. Alp. Res.* **35,** 129–143 (2003).
26. Matthews, J. A. "Little Ice Age" glacier variations in Jotunheim, southern Norway: a study in regionally controlled lichenometric dating of recessional moraines, with implications for climate and lichen growth rates. *Holocene* **15,** 1–19 (2005).
27. Food & Agriculture Organization. *Soil Map of the World* Vol. VIII, *North and Central Asia* (United Nations Educ. Cult. Org., 1978).
28. Jenkins, R. J. F. in *The Geological Record of Neoproterozoic Glaciations* (eds Arnaud, E., Halverson, G. P. and Shields-Zhou, G.) 693–698 (Geol. Soc. London Mem., 2011).
29. Ewing, S. A. *et al.* A threshold in soil formation at Earth's arid-hyperarid transition. *Geochim. Cosmochim. Acta* **70,** 5293–5322 (2006).
30. Grazhdankin, D. & Gerdes, H. Y. Ediacaran microbial colonies. *Lethaia* **40,** 201–210 (2007).

Supplementary Information is available in the online version of the paper.

Acknowledgements K. Lloyd, P. Coulthard, A. Coulthard, K. Anderson and D. Crawford facilitated permission to undertake research in Flinders Ranges National Park. B. Logan and M. Willison aided sampling of drill core at PIRSA, Glenside. T. Palmer and D. Atkins provided mathematical advice. Fieldwork was funded by the PRF fund of the American Chemical Society, and aided by C. Metzger and J. Gehling.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The author declares no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to G.J.R. (gregr@uoregon.edu).

LETTER

doi:10.1038/nature11664

Convergent acoustic field of view in echolocating bats

Lasse Jakobsen[1], John M. Ratcliffe[1] & Annemarie Surlykke[1]

Most echolocating bats exhibit a strong correlation between body size and the frequency of maximum energy in their echolocation calls (peak frequency), with smaller species using signals of higher frequency than larger ones[1,2]. Size–signal allometry or acoustic detection constraints imposed on wavelength by preferred prey size have been used to explain this relationship[1,3]. Here we propose the hypothesis that smaller bats emit higher frequencies to achieve directional sonar beams, and that variable beam width is critical for bats. Shorter wavelengths relative to the size of the emitter translate into more directional sound beams[4]. Therefore, bats that emit their calls through their mouths should show a relationship between mouth size and wavelength, driving smaller bats to signals of higher frequency. We found that in a flight room mimicking a closed habitat, six aerial hawking vespertilionid species (ranging in size from 4 to 21 g, ref. 5) produced sonar beams of extraordinarily similar shape and volume. Each species had a directivity index of 11 ± 1 dB (a half-amplitude angle of approximately 37°) and an on-axis sound level of 108 ± 4 dB sound pressure level referenced to 20 μPa root mean square at 10 cm. Thus all bats adapted their calls to achieve similar acoustic fields of view. We propose that the necessity for high directionality has been a key constraint on the evolution of echolocation, which explains the relationship between bat size and echolocation call frequency. Our results suggest that echolocation is a dynamic system that allows different species, regardless of their body size, to converge on optimal fields of view in response to habitat and task.

For echolocating bats, peak frequency in echolocation calls is negatively related to body size, a trend attributed to allometry[1,2]. However, similarly sized birds, anurans and most mammals use much lower frequencies for communication[6,7], so allometry does not adequately explain why bats use such high frequencies for echolocation. Indeed, atmospheric attenuation[8] increases rapidly with frequency, and echolocation range is much shorter than it would be at lower frequencies[9]. Thus, there must be some functional explanation why bats, especially smaller species, use such high-frequency sonar signals.

One hypothesis proposes that because small bats hunt small prey they use wavelengths short enough to be effectively reflected from their smallest quarry. However, this does not adequately explain bats' high-frequency calls, as it presumes that strong echoes will only be reflected from objects with diameters equal to or greater than the wavelength, which is not true. To reflect sound efficiently, the radius, a, of an object relative to the wavelength, λ, of the impinging sound has to fulfil $2\pi a/\lambda > 1$. Thus the effective diameter of the object only has to be greater than approximately $\lambda/3$ (ref. 10). Consequently, a 6-mm-diameter insect will reflect echoes efficiently at 20 kHz, which agrees well with data showing that even small insects (4–5 mm) reflect strong echoes down to 20 kHz[11]. Diet analysis also reveals that many bats take prey with wing lengths much shorter than the wavelength of their sonar calls[3].

We propose a new hypothesis to explain bat size–signal allometry. Specifically, that smaller bats are constrained to higher frequencies to achieve a sufficiently directional beam, because sound beams broaden with decreasing emitter size (Fig. 1). A directional sonar beam is critical for echolocators, focusing energy in a forward-directed cone and thus minimizing off-axis echoes and increasing on-axis intensity and therefore range[4,6,12].

For visual vertebrates, fields of view (FOV) depends on skull design and is relatively fixed[13]. For echolocators, acoustic FOV is the volume ensonified sufficiently to generate detectable echoes[12,14,15]. From FOV bats build and update their auditory world[16]. Bats can adapt FOV (beam range and width) dynamically by changing: (1) call energy (range); (2) call frequency by laryngeal control (width); and (3) emitter size by gape control (width). Hence, the acoustic FOV of most bats should be more flexible than visual FOV and directionality an integral component of echolocation call design. In closed habitat, where range is not crucial, a broader beam provides a wider FOV, enhancing peripheral object detection. We predict that morphologically similar bats orienting in the same habitat will produce sonar signals of uniform beam shape, converging on an optimal FOV.

Vespertilionidae comprises one-third of extant bat species and exhibits pronounced negative signal frequency to body size scaling[1]. We tested our predictions using six vespertilionid species of similar face and ear morphology, *Pipistrellus pygmaeus, Myotis daubentonii*, *Vespertilio murinus*, *Myotis dasycneme*, *Eptesicus serotinus* and *Nyctalus noctula*, that range in size from 4 to 21 g and produce calls over open field with peak frequencies of 20–55 kHz (Supplementary Table 1). We used a

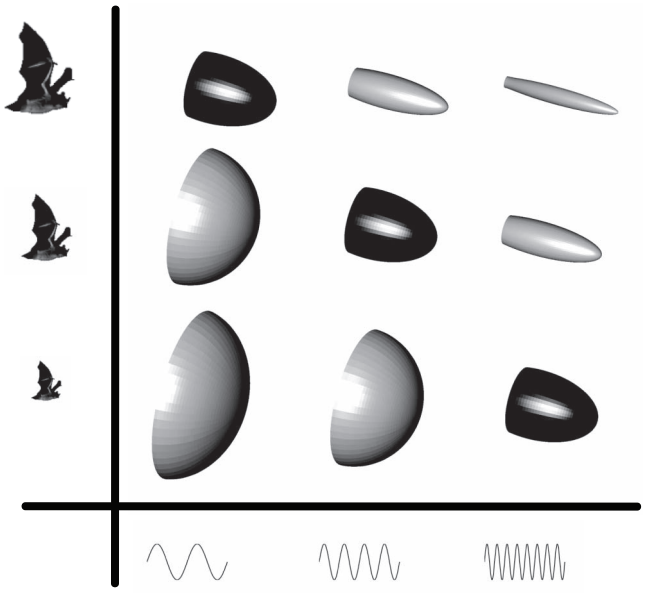


Figure 1 | Sonar beam width decreases as emitter (bat) size increases relative to wavelength. The beam width cartoons (three-dimensional figures) illustrate that for constant energy and emitter size, an increase in frequency (left to right), that is decrease in wavelength, focuses the energy in a sonar beam, to become narrower but longer, which at short distances counteracts the decrease in sonar range due to increased atmospheric attenuation at higher frequencies. The smaller the bat, the smaller is its emitter (mouth) size, and thus the broader its beam for constant frequency. Hence the smaller the bat, the higher the frequency required to maintain directionality of the biosonar beam.

[1]Sound Communication Group, Institute of Biology, University of Southern Denmark, DK-5230 Odense M, Denmark.

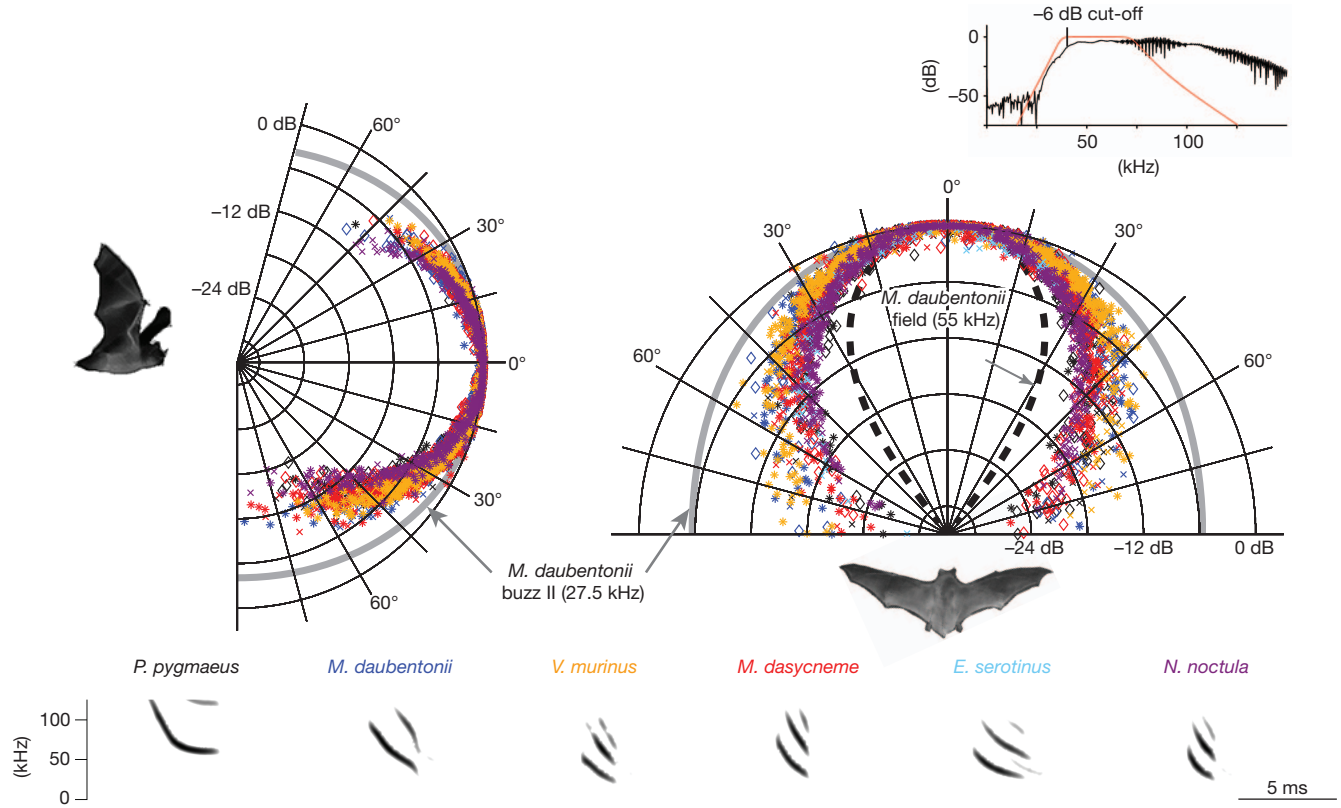


Figure 2 | Vertical and horizontal directionality in six species of vespertilionid bats. Vertical (left) and horizontal (right) directionality for *P. pygmaeus* (half-amplitude angle (HA) = 36°), *M. daubentonii* (HA = 40°), *V. murinus* (HA = 37°), *M. dasycneme* (HA = 40°), *E. serotinus* (HA = 34°) and *N. noctula* (HA = 34°). Marker colour corresponds to font colour of each species' name. Dashed black line indicates the much narrower directionality of *M. daubentonii* when flying in open field, whereas grey lines in both plots indicate the broad directionality *M. daubentonii* uses in the terminal phase of the prey pursuit. The bottom panel shows spectrograms of the calls emitted in the flight room by the six species in order of size from left to right. The inset in the top right corner shows a spectrum of a *M. daubentonii* call indicating the one octave band-pass filter starting at the −6 dB low-frequency cut-off.

multi-microphone array to determine sonar beam directionality as bats oriented in a large, screened flight room, and estimated their acoustic emitter size using the piston model[4]:

$$R_p(\theta) = \left| \frac{2 \times J_1(k \times a \times \sin(\theta))}{k \times a \times \sin(\theta)} \right|$$

where $R_p(\theta)$ is the ratio between the pressure on-axis and at a given angle θ, J_1 is a first-order Bessel function of the first kind, $k = 2\pi/\lambda$, λ is wavelength, and a is piston radius.

We measured directionality over a full octave band starting at −6 dB down from peak (Fig. 2). All bats emitted calls with similar beam widths. Half-amplitude angles were 37 ± 3°(Fig. 2). Call intensities also converged across species, 108 ± 4 dB root mean square sound pressure level at 10 cm (Supplementary Table 1). To quantify measured beam patterns we computed the directivity index of each call (Fig. 3). The directivity index (DI) compares on-axis sound pressure with the sound pressure of an omnidirectional emitter producing a signal of equal energy. Calculated DI values (10.7–12.1 dB) confirm directionality as nearly identical across species despite differences in emitter size and frequency. Maximum inter-specific DI differences of 1.4 dB (Fig. 3) are negligible under ecologically relevant conditions.

These similar DI values are striking because the range of DI values available to individual vespertilionid bats is greater than the inter-specific variation we observed in the flight room. *M. daubentonii* emits a narrower beam, with a DI of 16 dB, by opening its mouth wider over open field[12]. In the last phase of an aerial attack *M. daubentonii* and *E. serotinus* lower their call frequency an octave, emitting broader beams (DI of 6 dB)[15] (Fig. 3). Our data indicate that vespertilionids actively control directionality, adjusting emitter size and frequency, to converge on the same beam width. In closed habitats like our flight room, sonar range is not an issue and a broader beam provides peripheral information optimal to this habitat and task. Conversely, increasing range takes precedence while over open field, where, as demonstrated for *M.*

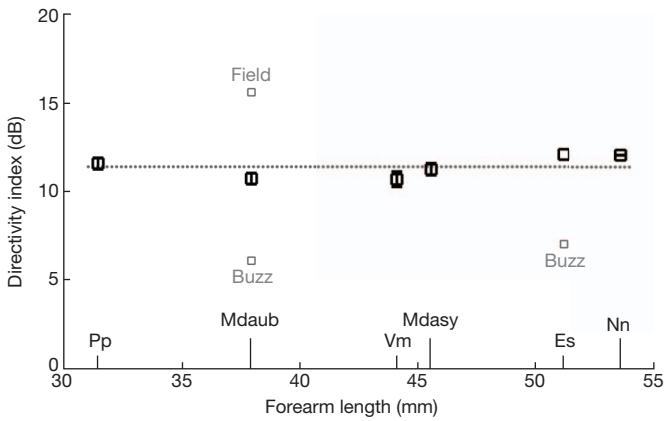


Figure 3 | Directivity index for six species of vespertilionid bats with forearm lengths ranging from 32 to 54 mm. Directivity indices (DI) (± s.d.) calculated from best fit of the piston model for total energy of the first harmonic (1 octave filter, see Fig. 2) of each species' echolocation call ($n = 1-3$ per species). DI compares the on-axis sound pressure with that of an omnidirectional emitter producing a signal of the same total energy that is DI = 0 dB. The larger the DI, the higher the directionality. *M. daubentonii* emits more directional search calls in the field with DI of 16 dB ('field'). *M. daubentonii* and *E. serotinus* emit broader beams with directivity indices of 6 dB in the last phase of the pursuit, the buzz II ('buzz'). The dotted line represents the mean of the DI values.

daubentonii[12], vespertilionid bats should emit more directional beams. Indeed, larger bats emit lower peak frequencies in open field, increasing range; in closed habitat, bandwidth is increased to include higher frequencies, enhancing resolution and temporal accuracy[17]. The two smallest species, particularly *P. pygmaeus*, increased peak frequency only slightly in the flight room, suggesting that, when orienting, smaller species regulate directionality almost exclusively by adjusting emitter size (Fig. 4).

To relate our acoustic models to actual species-specific mouth sizes (maximum gape) we measured upper and lower jaw length (craniomandibular joint to front teeth) and the ratio between distance from the craniomandibular joint to the origin (A) and insertion (B) of the superficial masseter (Fig. 4). The longer A is relative to B, the larger the maximal gape angle[18]. Assuming that inter-specific differences in A/B relate directly to differences in gape angle, we estimated gape height using a reported gape angle of 90° for *M. lucifugus*[19] and a measured A/B of 2.1. We used upper jaw width at the second incisors and third molars to estimate species-specific differences in maximum gape width (Supplementary Table 1).

To verify emitter size estimates from acoustic data, we compared them to actual measured data on mouth opening. Flying *M. daubentonii* had distances between the upper and lower lip of 5–8 mm (Supplementary Fig. 1), closely matching acoustic data (Fig. 4 and Supplementary Table 1). Subsequently, we compared emitter size estimates and morphometric measures and also compared these data to acoustic data at the species-level and using independent contrasts[20] (see Methods for further details).

Maximum gape height and width from skull measurements were comparable, but always larger than vertical and horizontal emitter size based on the piston model, suggesting that bats did not open their mouths to the maximum (Fig. 4). Gape estimates based on skull measures better predicted open space peak frequency than body mass

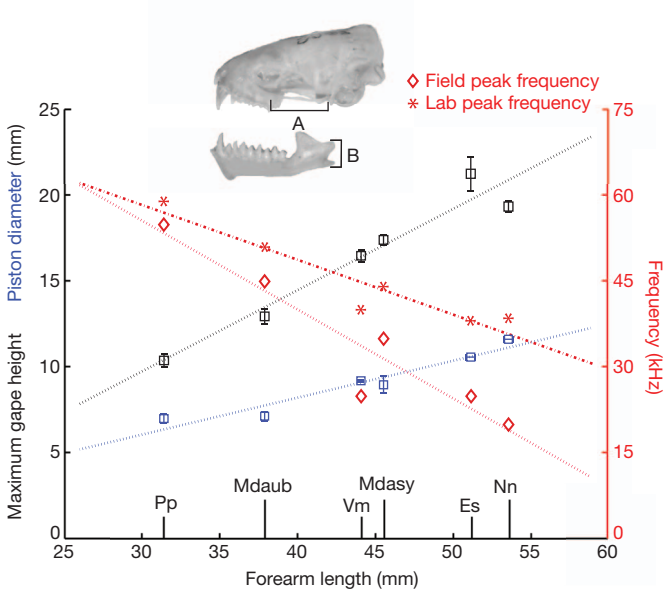


Figure 4 | Gape size estimated from skulls and from the piston model. From dry museum specimens we measured the distance from the craniomandibular joint to both the origin (A) and the insertion (B) of the superficial masseter muscle. The ratio between A and B relates directly to gape angle and was used together with upper and lower jaw length to estimate maximum vertical gape opening (black squares; see Methods). We also estimated gape diameter from the acoustic data using the piston model (blue squares). Gape size based on sonar sound beams were comparable, but always below maximum gape size based on skull metrics, indicating that the bats did not open their mouths to the maximum gape opening when flying in the flight room. Peak frequencies of echolocation calls in the flight room (red asterisks) were always higher than typical peak frequencies of the same species flying in the field (red diamonds).

(Supplementary Table 2). Similarly, emitter size estimates fitted well with second incisor distances (Supplementary Table 2). Forearm length correlated with open space peak frequency and maximum estimated gape height and width (Fig. 4 and Supplementary Table 2), as predicted by allometry and flight speed. Relationships between maximum estimated gape size and differences between open space and flight room frequencies also suggest that larger bats were using less than maximum gape in the flight room (Fig. 4 and Supplementary Tables 1 and 2).

We found that when orienting in the same context, six species of vespertilionid bats produce sonar beams with ~37° half-amplitude angles despite species-specific differences in frequency and emitter size. Over open field, larger, faster flying bats require longer detection ranges than smaller slower species to equally sample the distance they travel between calls (Supplementary Table 1). Greater range favours lower frequencies because of reduced transmission loss through atmospheric attenuation[9]. If larger bats lower peak frequency proportionally more than smaller bats, they must open their mouths wider to maintain beam width.

We propose that the requirement for a directional beam has driven the high frequencies of bat echolocation calls. Narrowing the beam focuses its energy and partially compensates for increased attenuation at higher frequencies, such that short ranges are not decreased for a given energy output (Supplementary Fig. 2). In the field, flight speed and call duration are proportional to bat size[21–23]; in our flight room all species flew at roughly the same speed (relative to the range for open field) and used similar call intensities and durations, with no clear correlation to bat size (Supplementary Table 1). Taking into account features defining the spatial filter formed by the sonar beam (call duration, intensity, beam width and flight speed), the six species sampled almost identical volumes. For open space, flight speed was positively related to intensity and negatively related to frequency (Supplementary Table 2) suggesting FOV is adjusted such that larger, faster bats monitor greater distances than smaller, slower bats.

Our results support our prediction that sonar beam width and acoustic FOV is dynamically controlled to best monitor a particular environment. We believe that the high frequencies emitted by bats are largely dictated by FOV, which acts as an evolutionary constraint on echolocation call design[24]. The receiving side, that is directional hearing, will also influence the bat's auditory perception[25], but the fact that bats adjust the beam to be narrower in open field[12] and broader in the last phase of pursuit[15] demonstrates that directionality on the receiving side does not diminish the importance of the outgoing acoustic FOV. Smaller bats typically have shorter, more gracile jaws and skulls and therefore smaller maximum emitter sizes. This forces smaller bats to use high echolocation call frequencies not because of preferred prey size or body size per se, but to obtain a directional beam over open field. The lack of correlation between bat size and signal frequency in phyllostomid bats[1] is consistent with such a view. Phyllostomids are nostril-emitting echolocators and sonar beam width depends on noseleaf dimensions, which do not scale with body size. Consequently, phyllostomids are not under the constraints resulting in size-frequency scaling in mouth-emitting species. Overall, our results depict bat echolocation as not only a viable substitute for vision under conditions of uncertain lighting, but as having unique advantages owing to flexible control of field of view.

METHODS SUMMARY

We recorded one *Eptesicus serotinus*, three *Myotis dasycneme*, three *M. daubentonii*, two *Nyctalus noctula*, three *Pipistrellus pygmaeus* and two *Vespertilio murinus* as they oriented in a 7 × 4.8 × 2.5 m flight room. Calls were recorded with twelve 0.25 inch 40BF G.R.A.S. microphones, amplified by 12 AA G.R.A.S. amplifiers, and sampled at 350 kHz per channel by an Avisoft 1216 Ultrasound Gate. Seven microphones were positioned horizontally 50 cm apart and five were positioned vertically 40 cm apart (two above and three below the central horizontal microphone).

We localized bats at each call using microphone arrival-time differences, compensating for transmission loss (spherical-spreading loss, atmospheric attenuation[8]

and angle on microphones[26]) by filtering the recorded signal by the impulse response of transmission loss. We filtered each call using a one-octave sixth order band-pass filter (ANSI-S1.1-1986-Standard) starting at the −6 dB low-frequency cut-off (Fig. 2 and Supplementary Table 1), which weighs directionality according to call energy distribution. We also calculated maximum beam width at minimum frequency (−6 dB low-frequency cut-off) and found similar directionality indices (8.9–10.4 dB). Root mean squared pressure of each compensated signal calculated for 95% energy content of 7.5 ms inclusive segments.

Beam aim was calculated in azimuth and elevation, using calls on-axis with the centre microphone. We fitted the piston model to emission pattern extracting equivalent piston radius, a, for each call.

We took gape heights for two flying *M. daubentonii* from the open-access repository http://www.ChiRoPing.org/data/index. Bats were video recorded at 500 frames per second and vertical gape heights measured. Eight flights per individual were analysed (Supplementary Fig. 1).

Skull photos ($n = 4$–7 per species) were exported to Image J v.1.38x, measurements made as described in ref. 18. We estimated gape height as:

$$Gape\ height_{\mathrm{species}} = \sqrt{a^2 + b^2 - 2 \times a \times b \times \cos\left(90^\circ \times \frac{\frac{A}{B_{\mathrm{species}}}}{\frac{A}{B_{\mathrm{lucifugus}}}}\right)}$$

a is upper jaw length, b lower jaw length, and A/B the ratio between distance from craniomandibular joint to origin and insertion of the superficial masseter (Fig. 4). Forearm measurements were taken from wet specimens ($n = 10$ per species).

Phylogenetically independent contrasts were generated using combined molecular phylogenies[27–29], actual branch lengths and the Crunch procedure (CAIC v.2.6.9, ref. 20).

Received 7 June; accepted 9 October 2012.
Published online 21 November 2012.

1. Jones, G. Scaling of echolocation call parameters in bats. *J. Exp. Biol.* **202,** 3359–3367 (1999).
2. Barclay, R. M. R. & Brigham, R. M. Prey detection, dietary niche breadth, and body size in bats: why are aerial insectivorous bats so small? *Am. Nat.* **137,** 693–703 (1991).
3. Houston, R. D., Boonman, A. M. & Jones, G. Do echolocation signal parameters restrict bats' choice of prey? In *Echolocation in Bats and Dolphins* (eds Thomas, J. A., Moss, C. F. & Vater, M) 339–345 (Chicago Univ. Press, 2004).
4. Mogensen, F. & Møhl, B. Sound radiation patterns in the frequency domain of cries from a vespertilionid bat. *J. Comp. Physiol.* **134,** 165–171 (1979).
5. Dietz, C., von Helversen, O. & Nill, D. *Handbuch der Fledermäuse Europas und Nordwestafrikas: Biologie, Kennzeichen, Gefährdung* 1st edn (Kosmos, 2007).
6. Bradbury, J. W. & Vehrencamp, S. L. *Principles of Animal Communication* 2nd edn (Sinauer Associates, 2011).
7. Suthers, R. A. Vocal mechanisms in birds and bats: a comparative view. *An. Acad. Bras. Cienc.* **76,** 247–252 (2004).
8. ANSI standard S1. *Method for the Calculation of the Absorption of Sound by the Atmosphere.* Report No. ANSI S1 26–1995 (ASA 113–1995) Acoustic. Soc. Am. (1995).
9. Lawrence, B. D. & Simmons, J. A. Measurements of atmospheric attenuation at ultrasonic frequencies and the significance for echolocation by bats. *J. Acoust. Soc. Am.* **71,** 585–590 (1982).
10. Pye, J. D. Is fidelity futile? The 'true' signal is illusory, especially with ultrasound. *Bioacoustics* **4,** 271–286 (1993).
11. Waters, D. A., Rydell, J. & Jones, G. Echolocation call design and limits on prey size: a case study using the aerial-hawking bat *Nyctalus leisleri. Behav. Ecol. Sociobiol.* **37,** 321–328 (1995).
12. Surlykke, A., Pedersen, S. B. & Jakobsen, L. Echolocating bats emit a highly directional sonar sound beam in the field. *Proc. R. Soc. B* **276,** 853–860 (2009).
13. Land, M. F. & Nilsson, D. E. *Animal Eyes* 2nd edn (Oxford Univ. Press, 2012).
14. Ghose, K. & Moss, C. F. Steering by hearing: a bat's acoustic gaze is linked to its flight motor output by a delayed, adaptive linear law. *J. Neurosci.* **26,** 1704–1710 (2006).
15. Jakobsen, L. & Surlykke, A. Vespertilionid bats control the width of their biosonar sound beam dynamically during prey pursuit. *Proc. Natl Acad. Sci. USA* **107,** 13930–13935 (2010).
16. Moss, C. F., Chiu, C. & Surlykke, A. Adaptive vocal behavior drives perception by echolocation in bats. *Curr. Opin. Neurobiol.* **21,** 645–652 (2011).
17. Schnitzler, H.-U. & Kalko, E. K. V. Echolocation by insect-eating bats. *Bioscience* **51,** 557–569 (2001).
18. Herring, S. W. & Herring, S. E. The superficial masseter and gape in mammals. *Am. Nat.* **108,** 561–576 (1974).
19. Kallen, F. C. & Gans, C. Mastication in little brown bat, *Myotis lucifugus. J. Morphol.* **136,** 385–420 (1972).
20. Purvis, A. & Rambaut, A. Comparative analysis by independent contrasts (CAIC): an Apple Macintosh application for analysing comparative data. *Comp. Appl. Biosci.* **11,** 247–251 (1995).
21. Norberg, U. M. & Rayner, J. M. V. Ecological morphology and flight in bats (Mammalia; Chiroptera): wing adaptations, flight performance, foraging strategy and echolocation. *Phil. Trans. R. Soc. B* **316,** 335–427 (1987).
22. Simmons, J. A., Fenton, M. B. & O'Farrell, M. J. Echolocation and pursuit of prey by bats. *Science* **203,** 16–21 (1979).
23. Baagøe, H. J. in *Recent Advances in the Study of Bats* (eds Fenton, M. B., Racey, P. & Rayner, J. M. V.) 57–74 (Cambridge Univ. Press, 1987).
24. Jones, G. & Teeling, E. C. The evolution of echolocation in bats. *Trends Ecol. Evol.* **21,** 149–156 (2006).
25. Wotton, J. M., Jenison, R. L. & Hartley, D. J. The combination of echolocation emission and ear reception enhances directional spectral cues of the big brown bat, *Eptesicus fuscus. J. Acoust. Soc. Am.* **101,** 1723–1733 (1997).
26. Brüel & Kjær. *Condenser Microphones and Microphone Preamplifiers for Acoustic Measurements. Data Handbook* (Brüel & Kjær, 1982).
27. Ruedi, M. & Mayer, F. Molecular systematics of bats of the genus *Myotis* (Vespertilionidae) suggests deterministic ecomorphological convergences. *Mol. Phylogenet. Evol.* **21,** 436–448 (2001).
28. Hoofer, S. R. & Van den Bussche, R. A. Molecular phylogenetics of the chiropteran family Vespertilionidae. *Acta Chiropterol.* **5,** 1–63 (2003).
29. Hulva, P., Horácek, I., Strelkov, P. P. & Benda, P. Molecular architecture of *Pipistrellus pipistrellus/Pipistrellus pygmaeus* complex (Chiroptera: Vespertilionidae): further cryptic species and Mediterranean origin of the divergence. *Mol. Phylogenet. Evol.* **32,** 1023–1035 (2004).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank H. Baagøe and B. Fenton for access to the bat collections at the Natural History Museum of Denmark and the Royal Ontario Museum, respectively, and R. Fisher and J. Hallam for providing gape heights of bats in flight. S. Brinkløv, C. Elemans, B. Falk, B. Fenton, J. Galef, P. Madsen, C. Moss, H.-U. Schnitzler and M. Wahlberg provided detailed comments that improved the manuscript. This study was funded by the Danish Council for Natural Sciences (FNU), Carlsberg, Oticon and the European Commission via the Seventh Framework Programme project ChiRoPing, Information Society Technologies Contract 215370. Animal capture and experimentation was approved by Skov- og Naturstyrelsen (Denmark).

Author Contributions L.J. was responsible for conducting the experiments and programming. All authors contributed to data analyses and the writing of the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to A.S. (ams@biology.sdu.dk).

LETTER

doi:10.1038/nature11665

Inhibition dominates sensory responses in the awake cortex

Bilal Haider[1], Michael Häusser[2] & Matteo Carandini[1]

The activity of the cerebral cortex is thought to depend on the precise relationship between synaptic excitation and inhibition[1–4]. In the visual cortex, in particular, intracellular measurements have related response selectivity to coordinated increases in excitation and inhibition[5–9]. These measurements, however, have all been made during anaesthesia, which strongly influences cortical state[10] and therefore sensory processing[7,11–15]. The synaptic activity that is evoked by visual stimulation during wakefulness is unknown. Here we measured visually evoked responses—and the underlying synaptic conductances—in the visual cortex of anaesthetized and awake mice. Under anaesthesia, responses could be elicited from a large region of visual space[16] and were prolonged. During wakefulness, responses were more spatially selective and much briefer. Whole-cell patch-clamp recordings of synaptic conductances[5,17] showed a difference in synaptic inhibition between the two conditions. Under anaesthesia, inhibition tracked excitation in amplitude and spatial selectivity. By contrast, during wakefulness, inhibition was much stronger than excitation and had extremely broad spatial selectivity. We conclude that during wakefulness, cortical responses to visual stimulation are dominated by synaptic inhibition, restricting the spatial spread and temporal persistence of neural activity. These results provide a direct glimpse of synaptic mechanisms that control sensory responses in the awake cortex.

To investigate how wakefulness affects the synaptic basis of visual selectivity, we made local field potential (LFP) recordings and whole-cell recordings of membrane potential (V_m) in layer 2/3 of the primary visual cortex (V1) in both anaesthetized and awake mice.

We first examined spontaneous activity and found that this activity was markedly affected by wakefulness (Fig. 1a, b). Under two widely used anaesthetic regimes, slow fluctuations in both V_m and LFP[18] were common (Fig. 1a and Supplementary Fig. 4). During wakefulness, these slow fluctuations were abolished and were replaced by higher frequency activity of both V_m and LFP (Fig. 1b and Supplementary Fig. 1). In nearly all cases, V_m was distributed unimodally during waking[19,20] but bimodally during anaesthesia (Supplementary Fig. 1a, b). Spontaneous firing rates were similarly low in the two conditions (anaesthetized, 0.3 ± 0.2 spikes s^{-1}, $n = 14$; awake, 0.1 ± 0.1 spikes s^{-1}, $n = 14$; $P = 0.07$). These results indicate that in awake mice, V1 rarely shows the spontaneous fluctuations that are common during anaesthesia or sleep[7,20] and that have been reported in area S1 of quietly awake mice[11].

We next probed visual responses with flashed bars and found that wakefulness had a striking effect on response duration (Fig. 1c–f). Briefly flashed bars (100 ms duration, at 1.5 s intervals) elicited long-lasting LFP responses under anaesthesia (Fig. 1c; 553 ± 22 ms, $n = 7$ mice) and much briefer responses during wakefulness (Fig. 1d; 171 ± 11 ms, $n = 7$; $P < 0.001$). This striking difference in LFP response duration was observed across the depth of the cortex (Supplementary Fig. 2). Awake V_m responses were also rapidly truncated (Fig. 1h; 148 ± 31 ms, $n = 14$) compared with anaesthetized responses (Fig. 1g; 553 ± 73 ms, $n = 14$; and Supplementary Fig. 3). The prolonged responses in both V_m and LFP were remarkably similar across anaesthetic regimes and persisted regardless of the depth of anaesthesia (Supplementary Figs 4 and 5).

This marked difference in awake and anaesthetized responses was not confounded by spontaneous alternations of excitability that are present during anaesthesia (Fig. 2). We asked whether responses under anaesthesia differed when neurons were spontaneously hyperpolarized (down) or depolarized (up)[7]. After correcting for the tendency of V_m to spontaneously alternate between these two states (Fig. 2c), we found that during anaesthesia, the V_m responses evoked from either state (hyperpolarized or depolarized) were remarkably similar in amplitude and duration (Fig. 2e) and were much longer than responses during wakefulness (Fig. 2f).

Responses in awake mice were more selective across visual space than responses under anaesthesia (Fig. 3a–e). V_m responses were twice as spatially selective during wakefulness as under anaesthesia. This difference in spatial localization was even more pronounced for spikes[21] (Fig. 3e), even when we accounted for sustained responses during anaesthesia by restricting the spike counts to the earliest portion of the sensory response (0–200 ms; Supplementary Fig. 6).

We observed fewer visually evoked spikes during waking than during anaesthesia (Fig. 3b). This difference was particularly evident during the stimulation of regions that surround the centre of the receptive field. Under anaesthesia, stimuli in these regions evoked significantly more firing (0.6 ± 0.2 spikes per trial) than did blank stimuli (0.3 ± 0.2 spikes per trial; $P < 0.001$). By contrast, in awake animals, stimuli in this region produced no net increase in spikes above spontaneous activity (0.08 ± 0.03 for surround stimuli versus 0.08 ± 0.04 for blank stimuli; $P = 0.89$).

The lower spike counts observed during wakefulness were also associated with a reduced variability in V_m responses (awake s.d., 3.5 ± 0.03 mV, $n = 14$; anaesthetized s.d., 4.4 ± 0.03 mV, $n = 14$; $P < 0.001$). This lower variability reduced the number of threshold crossings of V_m in awake animals, despite peak responses that were, on average, more depolarized than those during anaesthesia (Fig. 3d). The spike threshold did not differ between wakefulness (-40.0 ± 1.1 mV) and anaesthesia (-40.0 ± 1.2 mV; $P = 0.6$). However, spikes evoked during wakefulness were quickly followed by a significant and long-lasting hyperpolarization (Fig. 3f). We hypothesized that this hyperpolarization—and indeed the finer spatiotemporal resolution of the awake responses—was indicative of enhanced synaptic inhibition.

To test this hypothesis, we blocked the intrinsic conductances[22,23] in single neurons and recorded synaptic currents in voltage-clamp mode near the reversal potentials for glutamate-mediated excitation and GABA (γ-aminobutyric acid)-mediated inhibition (Supplementary Fig. 7). We then estimated the relative change in total conductance (ΔG; Supplementary Figs 7 and 8) visible at the soma[24] in response to visual stimulation. Resting conductance and peak-evoked conductance were unaffected by wakefulness (Supplementary Fig. 7).

[1]UCL Institute of Ophthalmology, University College London, 11–43 Bath Street, London EC1V 9EL, UK. [2]Wolfson Institute for Biomedical Research and Department of Neuroscience, Physiology and Pharmacology, University College London, Gower Street, London WC1E 6BT, UK.

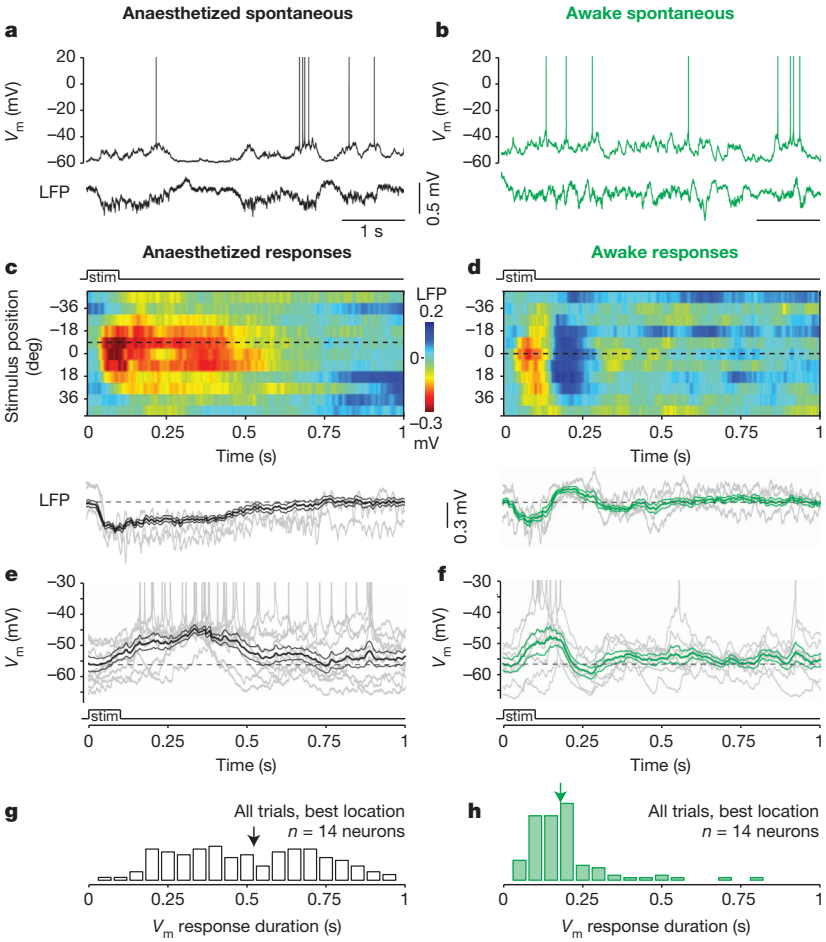


Figure 1 | Spontaneous and evoked activity in the anaesthetized and awake visual cortex (V1). **a**, V_m and simultaneous LFP measured in V1 under anaesthesia. The spikes are shown truncated at +20 mV. **b**, V_m and simultaneous LFP measured in V1 in awake animals. **c**, **d**, Visually evoked LFP responses across space during anaesthesia (**c**) and wakefulness (**d**). An average of 15 trials per location was used. Top, the dashed line indicates the best location. Bottom, single trial responses (grey) and the average (mean ± s.e.m.) response (black or green) with the stimulus (stim) at the best location. deg, degrees. **e**, V_m responses to stimuli at the best location while under anaesthesia: single trials (grey) and mean (± s.e.m.; black). The spikes are shown truncated at −30 mV. **f**, V_m responses to stimuli at the best location during wakefulness. **c**–**f**, $n = 4$ different mice. **g**, **h**, Normalized probability distributions of V_m response durations to stimuli at the best location, across the population ($n = 14$), measured under anaesthesia (**g**) and during wakefulness (**h**). The arrow indicates the mean duration.

Therefore, the differences between awake and anaesthetized V_m must result from changes in the relative strength of the excitatory conductance (ΔG_e) and inhibitory conductance (ΔG_i).

Under anaesthesia, the estimated excitatory conductance and inhibitory conductance behaved as expected from previous studies[1,5,7,8,17,22,23,25,26] (Fig. 4a). On stimulation, increased excitation was quickly followed by inhibition. Then, excitation and inhibition co-varied at a sustained and elevated level for hundreds of milliseconds after the stimulus offset. In other words, excitation and inhibition were balanced in that they had a similar amplitude and time course.

During wakefulness, by contrast, inhibition dominated excitation during the entire time course of the visual response (Fig. 4b). During

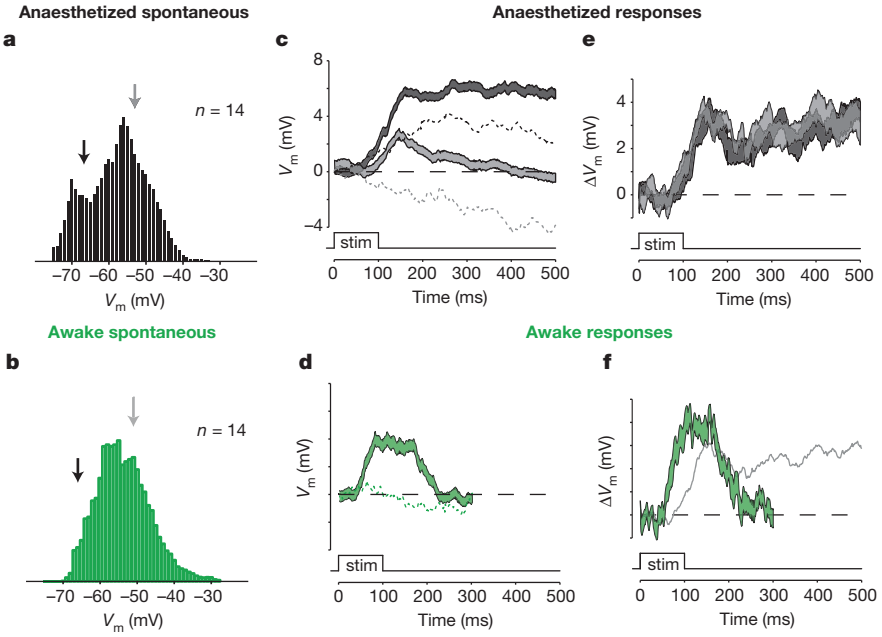


Figure 2 | Anaesthetized responses are long lasting regardless of cortical state. **a**, **b**, Normalized probability distributions of spontaneous V_m during anaesthesia ($n = 14$ neurons) (**a**) and wakefulness ($n = 14$ neurons) (**b**). **c**, Mean (± s.e.m.) anaesthetized V_m responses (solid lines), sorted by pre-stimulus V_m level. Depolarized (grey) and hyperpolarized (black) groups are shown (with the mean V_m of the two groups indicated by arrows of the corresponding colour in **a**). The pre-stimulus baseline V_m was subtracted before averaging. The dashed lines indicate the average spontaneous V_m (in response to blank stimuli), sorted similarly into two groups ($n = 14$ neurons). **d**, As for **c**, during wakefulness ($n = 14$). **e**, Average anaesthetized V_m response for hyperpolarized (black) and depolarized (grey) trials, after subtraction of spontaneous V_m traces. **f**, Average awake V_m response for all trials, after subtraction of spontaneous V_m traces. The grey line shows the average of all of the anaesthetized responses (for comparison).

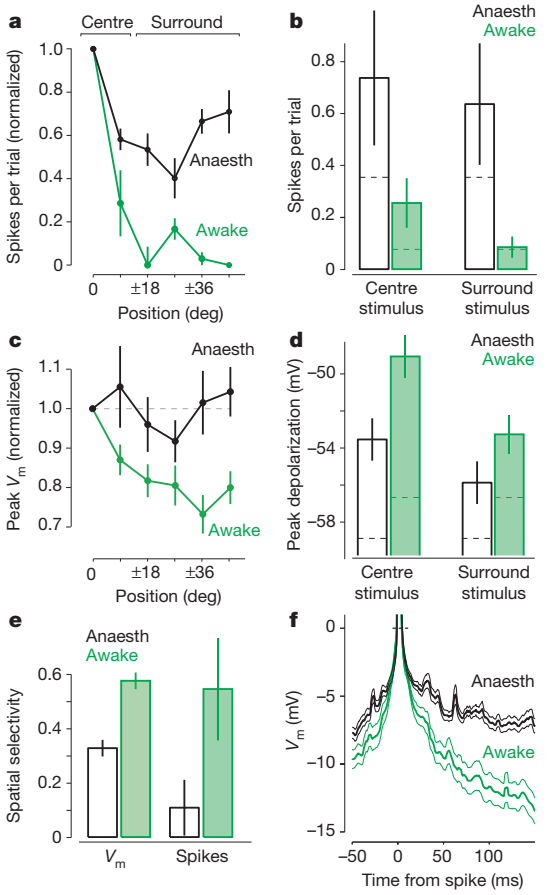


Figure 3 | Responses are spatiotemporally restricted during waking. **a**, The number of spikes evoked per trial (normalized to each neuron's response at the best location). Symmetrical locations on either side of 0° were combined. The centre is defined as 0° ± 9°, and the surround is defined as ±18° to ±45°. The response window is defined by the average duration of the population's V_m response (Fig. 1g, h) ($n = 14$ for each group). **b**, Under anaesthesia (anaesth, black), centre and surround stimuli evoked more spikes ($P < 0.001$ for both) than during spontaneous activity (dashed line). During wakefulness (green), there were fewer spikes than under anaesthesia ($P < 0.001$ for both stimulus locations); the centre stimuli evoked more spikes than did the surround stimuli ($P < 0.009$), and the surround stimuli did not evoke a significantly different response from spontaneous activity. Nine of 14 neurons were active during anaesthesia, and 5 of 14 were active during waking. **c**, As for **a**, for peak V_m responses (normalized to each neuron's response at the best location). **d**, As for **b**, for peak V_m responses. The responses to centre stimuli were greater than to surround stimuli in both anaesthetized and awake animals ($P < 0.04$ for both conditions), and all responses were greater than spontaneous activity (dashed lines, $P < 0.001$ for all). The awake responses were larger than the anaesthetized responses ($P < 0.001$ for both stimuli locations). **e**, The V_m and the spike responses were more spatially selective during waking (V_m, anaesthetized, 0.3 ± 0.1; awake, 0.6 ± 0.1; $P < 0.001$; and spikes, anaesthetized, 0.1 ± 0.1; awake, 0.6 ± 0.2; $P < 0.001$). **f**, The spike-triggered average of V_m under anaesthesia and during wakefulness. The spike threshold (the peak of the second derivative of V_m) was aligned at 0 mV. **a**–**f**, mean ± s.e.m.

the initial 100 ms of the response, the ratio of inhibition (ΔG_i) to excitation (ΔG_e) was 1.4–2.9 (interval defined by the geometric s.e.m. around the geometric mean, see Methods), which is significantly larger than the ratio measured under anaesthesia (0.7–1.0) ($P < 0.05$, one-tailed two-sample t-test; Supplementary Fig. 9). Inhibition remained above baseline significantly longer than did excitation (by 29 ± 13 ms; $P < 0.001$), but both excitation and inhibition disengaged within 200 ms of the stimulus, mirroring the rapid termination of awake V_m and LFP responses observed earlier

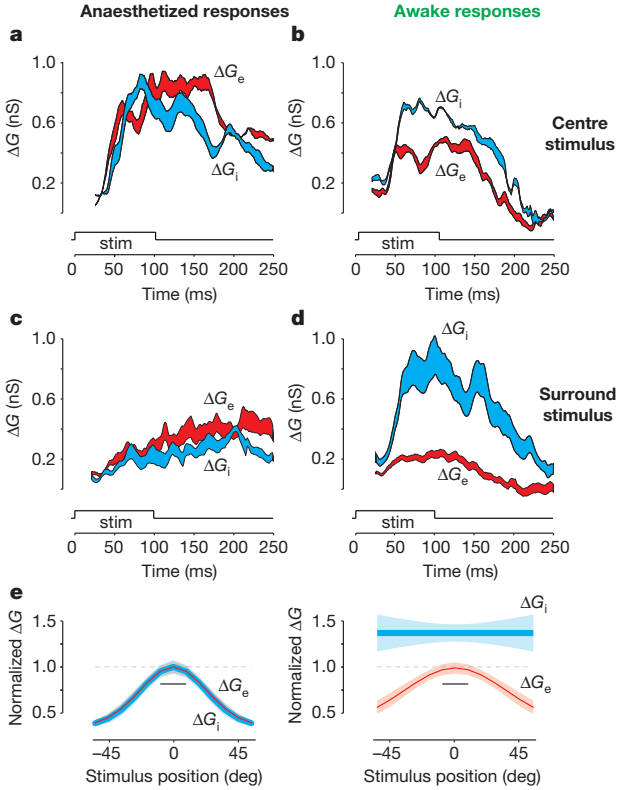


Figure 4 | Visually evoked conductance is dominated by inhibition in awake V1. **a**, **b**, ΔG_e (red) and ΔG_i (blue) evoked by centre stimulation during anaesthesia ($n = 5$) and waking ($n = 8$). **c**, ΔG_e and ΔG_i evoked by surround stimulation during anaesthesia (**c**) and waking (**d**) (for the same neurons as in **a** and **b**). **e**, Spatial profiles of excitation and inhibition under anaesthesia (left) and wakefulness (right). ΔG_e and ΔG_i were normalized to peak ΔG_e for centre stimuli (grey dashed line) for each neuron and then averaged across the population. Data were fitted with a Gaussian function (mean ± s.d., shaded) or a linear function (for the awake G_i; mean ± s.d.). Scale bar, 18° width across the receptive field centre. **a**–**d**, mean ± s.e.m.

(compare with Figs 1 and 2). Wakefulness also reduced the amplitude of ΔG_e, presumably because the intracortical sources of excitation were themselves subject to the same enhanced inhibition as the recorded neuron.

Strikingly, in awake animals, visually evoked inhibition was strongly activated even by stimuli that were placed far from the receptive field centre (Fig. 4d and Supplementary Fig. 9). During anaesthesia, by contrast, placing stimuli in these surrounding regions evoked little inhibition (Fig. 4c). Across all recordings in anaesthetized mice ($n = 6$), ΔG_i over the first 100 ms of the response was 1.8–2.4-fold larger for stimuli in the centre than in the surround (ΔG_i centre/surround ratio significantly >1; $P < 0.005$, one-tailed t-test). In awake mice, ΔG_i was remarkably unselective for position: it was not significantly different in amplitude regardless of whether it was elicited from the centre or from the surrounding regions ($P = 0.07$). In fact, in every neuron recorded during wakefulness ($n = 16$; Supplementary Fig. 9), the ΔG_i to ΔG_e ratio for surround stimulation was greater than 1; for all recorded neurons, the $\Delta G_i/\Delta G_e$ ratio evoked by surround stimuli was 0.8–1.1 under anaesthesia and a much larger 2.7–3.5 during wakefulness ($P < 0.01$).

Although retinotopy is the primary determinant of sensory responses in the visual cortex, synaptic conductances are also known to depend on stimulus orientation[5,8,22,23]. We asked whether changing the stimulus orientation would have an effect on the observations. The results were similar to those described above: regardless of the stimulus orientation during wakefulness, V_m responses were brief, and inhibition dominated across visual space (Supplementary Figs 10 and 11).

Taken together, these data identify a novel characteristic of awake cortical processing—elevated and spatially extended inhibition—that is associated with sensory responses that are more spatiotemporally selective (Fig. 4e). Previous measurements of sensory responses in anaesthetized animals have led to debate about the role of inhibition[1–5,7,8,17,22,23,25,26]; our findings show that inhibition is a decisive factor in the awake cortex: it dominates excitation in amplitude and over time (Fig. 4b, d) and is evoked from regions of visual space that extend far beyond the central regions of the receptive field (Fig. 4e). This finding of increased inhibition during wakefulness is consistent with earlier suggestions[11,27,28] and indicates a regime of sensory processing that cannot be observed during anaesthesia or sleep, in which more-balanced excitation and inhibition are evoked from large regions of space and persist long after the stimulus has disappeared. The increased inhibition in the awake cortex is ideally poised to extinguish any spatial or temporal spread of feedforward activity that is elicited by a sensory input. Accordingly, during wakefulness, we observed a brisk and highly selective impulse response to spatially localized visual stimuli.

At present, it is unclear which factors regulate the strength of inhibition in the awake cortex. Neuromodulators can desynchronize LFP and V_m[29], depolarize interneurons[30], and alter response reliability and sensory perception[10,13,19]. It will be important to examine such factors in this context, including the contributions of laminar connectivity and interneuron subtypes[1–4] to the increased inhibitory conductances that we observed during wakefulness. Having identified inhibition as a major determinant of spatially selective and temporally succinct visual responses in the awake cortex, we suggest that behavioural factors such as attention and reward may also exert their influence by modulating inhibition.

METHODS SUMMARY

All recordings were performed in layer 2/3 of monocular V1 (0.5 mm anterior and 2.0 mm lateral from lambda) in female C57BL/6J mice (4–6 weeks of age). Anaesthesia was induced with 10^{-5} mg chlorprothixene per kg body weight and 1.5 mg urethane per kg body weight (Figs 1–4) or with 10^{-5} mg chlorprothixene per kg body weight and 0.25–1% isoflurane in O_2. Awake mice were habituated to head fixation over 4–5 days. The LFP was recorded with pipettes filled with HEPES-buffered artificial cerebrospinal fluid (which consisted of 135 mM NaCl, 5.4 mM KCl, 5 mM HEPES, 1 mM $MgCl_2$ and 1.8 mM $CaCl_2$; pH 7.3). Patch pipettes (4–7 MΩ) were filled with standard internal solution (135 mM potassium gluconate, 6 mM KCl, 10 mM HEPES, 4 mM MgATP, 0.3 mM Na_2ATP, 0.1 mM EGTA and 8 mM phosphocreatine; pH adjusted to 7.3; 290–295 mOsM) for current-clamp recordings. For voltage-clamp recordings, a 140 mM caesium-methanesulphonate-based solution also included 0.5 mM QX-314 and 5 mM tetraethylammonium (TEA)[22,23]. V_m was not corrected for the junction potential. The series resistance was compensated online and was monitored throughout voltage-clamp recordings (anaesthetized mice, 15 ± 2 MΩ; and awake mice, 16 ± 2 MΩ). Conductance was calculated using the instantaneous current–voltage relationship at two holding potentials (near −80 and +20 mV), relative to baseline, as previously described[5,17,22]. Vertically oriented black and white bars (9° wide, 100% contrast, 100 ms duration and 1.5 s interstimulus interval) were presented monocularly, one at a time in randomly chosen positions (±45°). Spatial selectivity was defined as $(R_{Centre} - R_{Surround})/(R_{Centre} + R_{Surround})$, where R is the peak V_m (or number of spikes per trial) averaged across the centre or surround locations. Mean ± s.e.m. is reported throughout, unless noted. For ratios, we used the geometric mean and geometric s.e.m., defined as exp{mean[log(ratios)]} and exp{s.e.m.[log(ratios)]}. Statistical testing ($\alpha = 0.05$) was carried out using Wilcoxon signed-rank tests (paired data), rank-sum tests (unpaired data) and sign tests (difference from unity), unless noted.

Full Methods and any associated references are available in the online version of the paper.

Received 8 May; accepted 8 October 2012.
Published online 21 November 2012.

1. Isaacson, J. S. & Scanziani, M. How inhibition shapes cortical activity. *Neuron* **72,** 231–243 (2011).
2. Atallah, B. V., Bruns, W., Carandini, M. & Scanziani, M. Parvalbumin-expressing interneurons linearly transform cortical responses to visual stimuli. *Neuron* **73,** 159–170 (2012).
3. Wilson, N. R., Runyan, C. A., Wang, F. L. & Sur, M. Division and subtraction by distinct cortical inhibitory networks *in vivo*. *Nature* **488,** 343–348 (2012).
4. Lee, S. H. *et al.* Activation of specific interneurons improves V1 feature selectivity and visual perception. *Nature* **488,** 379–383 (2012).
5. Borg-Graham, L., Monier, C. & Frégnac, Y. Visual input evokes transient and strong shunting inhibition in visual cortical neurons. *Nature* **393,** 369–373 (1998).
6. Cardin, J. A., Kumbhani, R. D., Contreras, D. & Palmer, L. A. Cellular mechanisms of temporal sensitivity in visual cortex neurons. *J. Neurosci.* **30,** 3652–3662 (2010).
7. Haider, B. & McCormick, D. A. Rapid neocortical dynamics: cellular and network mechanisms. *Neuron* **62,** 171–189 (2009).
8. Mariño, J. *et al.* Invariant computations in local cortical networks with balanced excitation and inhibition. *Nature Neurosci.* **8,** 194–201 (2005).
9. Priebe, N. J. & Ferster, D. Inhibition, spike threshold, and stimulus selectivity in primary visual cortex. *Neuron* **57,** 482–497 (2008).
10. Harris, K. D. & Thiele, A. Cortical state and attention. *Nature Rev. Neurosci.* **12,** 509–523 (2011).
11. Crochet, S., Poulet, J. F., Kremer, Y. & Petersen, C. C. Synaptic mechanisms underlying sparse coding of active touch. *Neuron* **69,** 1160–1175 (2011).
12. Gilbert, C. D. & Sigman, M. Brain states: top-down influences in sensory processing. *Neuron* **54,** 677–696 (2007).
13. Goard, M. & Dan, Y. Basal forebrain activation enhances cortical coding of natural scenes. *Nature Neurosci.* **12,** 1444–1449 (2009).
14. Niell, C. M. & Stryker, M. P. Modulation of visual responses by behavioral state in mouse visual cortex. *Neuron* **65,** 472–479 (2010).
15. Wörgötter, F. *et al.* State-dependent receptive-field restructuring in the visual cortex. *Nature* **396,** 165–168 (1998).
16. Bringuier, V., Chavane, F., Glaeser, L. & Frégnac, Y. Horizontal propagation of visual activity in the synaptic integration field of area 17 neurons. *Science* **283,** 695–699 (1999).
17. Wehr, M. & Zador, A. M. Balanced inhibition underlies tuning and sharpens spike timing in auditory cortex. *Nature* **426,** 442–446 (2003).
18. Steriade, M., Nunez, A. & Amzica, F. A novel slow (< 1 Hz) oscillation of neocortical neurons *in vivo*: depolarizing and hyperpolarizing components. *J. Neurosci.* **13,** 3252–3265 (1993).
19. Constantinople, C. M. & Bruno, R. M. Effects and mechanisms of wakefulness on local cortical networks. *Neuron* **69,** 1061–1068 (2011).
20. Steriade, M., Timofeev, I. & Grenier, F. Natural waking and sleep states: a view from inside neocortical neurons. *J. Neurophysiol.* **85,** 1969–1985 (2001).
21. Simons, D. J., Carvell, G. E., Hershey, A. E. & Bryant, D. P. Responses of barrel cortex neurons in awake rats and effects of urethane anesthesia. *Exp. Brain Res.* **91,** 259–272 (1992).
22. Liu, B. H. *et al.* Broad inhibition sharpens orientation selectivity by expanding input dynamic range in mouse simple cells. *Neuron* **71,** 542–554 (2011).
23. Tan, A. Y., Brown, B. D., Scholl, B., Mohanty, D. & Priebe, N. J. Orientation selectivity of synaptic input to neurons in mouse and cat primary visual cortex. *J. Neurosci.* **31,** 12339–12350 (2011).
24. Williams, S. R. & Mitchell, S. J. Direct measurement of somatic voltage clamp errors in central neurons. *Nature Neurosci.* **11,** 790–798 (2008).
25. Haider, B. *et al.* Synaptic and network mechanisms of sparse and reliable visual cortical activity during nonclassical receptive field stimulation. *Neuron* **65,** 107–121 (2010).
26. Ozeki, H., Finn, I. M., Schaffer, E. S., Miller, K. D. & Ferster, D. Inhibitory stabilization of the cortical network underlies visual surround suppression. *Neuron* **62,** 578–592 (2009).
27. Rudolph, M., Pospischil, M., Timofeev, I. & Destexhe, A. Inhibition determines membrane potential dynamics and controls action potential generation in awake and sleeping cat cortex. *J. Neurosci.* **27,** 5280–5290 (2007).
28. Swadlow, H. A. Thalamocortical control of feed-forward inhibition in awake somatosensory 'barrel' cortex. *Phil. Trans. R. Soc. Lond. B* **357,** 1717–1727 (2002).
29. Steriade, M., Amzica, F. & Nunez, A. Cholinergic and noradrenergic modulation of the slow (approximately 0.3 Hz) oscillation in neocortical cells. *J. Neurophysiol.* **70,** 1385–1400 (1993).
30. Arroyo, S., Bennett, C., Aziz, D., Brown, S. P. & Hestrin, S. Prolonged disynaptic inhibition in the cortex mediated by slow, non-α7 nicotinic excitation of a specific subset of cortical interneurons. *J. Neurosci.* **32,** 3859–3864 (2012).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank T. Sato, A. Saleem and A. Ayaz for help with procedures; S. L. Smith, C. Schmidt-Hieber and K. Powell for advice on recordings, and A. Roth, M. Scanziani and D. McCormick for comments. We are grateful to the National Science Foundation, the European Research Council, the Wellcome Trust, the Medical Research Council and the Gatsby Charitable Foundation for financial support. M.C. holds the GlaxoSmithKline/Fight for Sight Chair in Visual Neuroscience.

Author Contributions B.H. performed the experiments. B.H. and M.C. performed the analyses. B.H., M.H. and M.C. designed the study and wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to B.H. (b.haider@ucl.ac.uk).

METHODS

Female C57BL/6J mice aged 4–6 weeks were used in both anaesthetized and awake recordings, and all procedures were performed under license from the UK Home Office in accordance with the Animal (Scientific Procedures) Act 1986.

Recordings under anaesthesia. Anaesthesia was induced with 10^{-5} mg chlorprothixene per kg body weight and 1.5 mg urethane per kg body weight (10% w/v dissolved in lactated Ringer's solution, administered intraperitoneally) (for all data presented in Figs 1–4) or with 10^{-5} mg chlorprothixene per kg body weight and 0.25–1% isoflurane in O_2. In addition, 0.07 mg atropine sulphate per kg body weight was injected subcutaneously to maintain clear airways; 2 mg dexamethasone per kg body weight was administered intramuscularly to prevent oedema; and 4 mg Rimadyl per kg body weight was administered subcutaneously as an analgesic and anti-inflammatory. Body temperature was maintained at 37.1 °C with a feedback-regulated heating pad. A tracheotomy was performed (during urethane experiments), and animals were passively ventilated with pure O_2. The stimulated eye was protected with a gas-permeable contact lens. The unstimulated eye was gently sutured shut. A small head plate with a chamber was implanted over monocular regions of the visual cortex (0.5 mm anterior and 2 mm lateral from the lambda suture). Two small (<0.5 mm) craniotomies were drilled (within 200–400 μm of each other) for the separate entry of LFP and patch pipettes. A third craniotomy was drilled over the anterior retrosplenial cortex for the insertion of an Ag/AgCl reference wire. The dura was carefully removed immediately before insertion of whole-cell pipettes. Warm HEPES-buffered artificial cerebrospinal fluid (aCSF, 135 mM NaCl, 5.4 mM KCl, 5 mM HEPES, 1 mM $MgCl_2$ and 1.8 mM $CaCl_2$; pH 7.3) filled the chamber to prevent desiccation and maintain ionic balance.

Recordings during wakefulness. Awake mice were implanted with a lightweight head plate with a chamber under isoflurane anaesthesia and were allowed to recover for 2 days. On the third day, animals were acclimated to head fixation for 10–60 min, with a sweet-liquid reward given at the start and end of training. The duration of training was increased over 4–5 days until mice were sitting comfortably for 3 h. On the day of recording, craniotomies, as described above, were performed under isoflurane anaesthesia. The animal was allowed to recover for 3 h, and a single 3 h recording session commenced. The dura was opened immediately before recording. Mice showed frequent grooming, whisking and postural adjustments during experiments. They were also given a sweet-liquid reward every 30 min to keep them comfortable and hydrated.

Recordings. Patch pipettes (4–7 MΩ) were pulled on a PC-10 pipette puller (Narishige) and filled with internal solution containing 135 mM potassium gluconate, 6 mM KCl, 10 mM HEPES, 4 mM MgATP, 0.3 mM Na_2ATP, 0.1 mM EGTA and 8 mM phosphocreatine (pH adjusted to 7.3 with KOH, 290–295 mOsM). For voltage-clamp recordings, potassium gluconate was substituted with 140 mM caesium methanesulphonate. QX-314 (0.5 mM) and tetraethylammonium (TEA) (5 mM) were included to block voltage-gated Na^+ and K^+ conductances. While searching for neurons, the diffusion of blockers was limited by using low pressures (20–30 mbar); more importantly, the LFP electrode served as an internal control for visual responsiveness. LFP response amplitudes and tuning were unaltered after using pipettes containing channel blockers.

Whole-cell recordings were obtained with a MultiClamp 700B amplifier in voltage-clamp mode using standard techniques[31]. Pipette capacitance was neutralized before break-in. The average seal resistances for anaesthetized and awake voltage-clamp recordings were 3.8 ± 1.7 GΩ and 6.3 ± 1.3 GΩ, respectively. The average seal resistances for anaesthetized and awake current-clamp recordings were 4.5 ± 1.2 GΩ and 5.3 ± 1.1 GΩ, respectively. The initial access resistance ranged from 9 MΩ to 30 MΩ. This value generally increased within a few minutes and then stabilized. If the initial access was greater than 50 MΩ, the recording was not included for further analysis. Access resistance during visual stimulation did not differ between anaesthetized and awake recordings in current-clamp experiments (anaesthetized mice, 36 ± 6 MΩ; and awake mice, 39 ± 3 MΩ) or voltage-clamp experiments (anaesthetized mice, 38 ± 5 MΩ; and awake mice, 38 ± 4 MΩ). Access resistance was compensated online during voltage-clamp recordings and was optimized manually. Series compensation averaged 63 ± 1%, yielding an effective series resistance of 15 ± 2 MΩ and 16 ± 2 MΩ during anaesthetized and awake voltage-clamp recordings, respectively. The junction potential was not corrected. The voltage division through the uncompensated series resistance was corrected offline[17].

This method limits the distortion of the somatically recorded current by ongoing voltage fluctuations and membrane capacitance[5,22]; however, because of the necessity for intrinsic current blockade, it also masks potential contributions from active dendritic conductances[32]. This method also requires multiple trials and holding potentials; it thus cannot be used to simultaneously assess the relationship between excitation and inhibition within single trials[33].

As in previous studies[2,34], to better nullify excitatory currents, we used a holding potential of about +20 mV rather than 0 mV, as this mitigates the voltage decay across the dendrites. Current-clamp recordings had 0 current injection during stimulation protocols. Input resistance and series resistance were monitored between protocols with current pulse trains or voltage steps. Firing rate adaptation to supra-threshold pulses and broad spike widths confirmed that our recordings were from regular-spiking pyramidal neurons[35]. Across all anaesthetized ($n = 22$; −305 ± 27 μm) and awake ($n = 30$; −279 ± 27 μm) whole-cell recordings, there was no significant difference in laminar depth ($P = 0.94$), as estimated from the micromanipulator reading after it was zeroed upon contact with the cortical surface.

Acquisition, visual stimulation and analysis. All analyses and acquisitions were performed in MATLAB. V_m was low-pass filtered at 20 kHz. During voltage-clamp experiments, membrane current (I_m) was low-pass filtered at 2 kHz. LFP was filtered from 0.1–100 Hz. Data were acquired using a National Instruments board and were synchronized to stimulus onset with a photodiode signal.

The liquid crystal display monitor was positioned 24 cm from the mouse and at 0° elevation and 30–45° azimuth. The azimuthal position was adjusted so that the centre of the LFP receptive field was roughly centred on the screen. Stimuli were presented on a grey background and randomly interleaved across space, with 8–20 repetitions per stimulus location. A blank screen (the grey background) was randomly presented every 11 stimulus presentations for the same duration as the stimulus trials. Stimuli were delivered monocularly. The unstimulated eye was gently sutured (anaesthetized mice) or shielded by black aluminium foil (awake mice). Throughout the entire recording session, the monitor was continuously illuminated with the same grey screen. This was to ensure that the spontaneous cortical state was not affected by transitions to and from complete darkness of the monitor.

In the awake recordings, on establishing a stable whole-cell configuration, 1–2 min of spontaneous activity was recorded before the presentation of any flashed bars. These data were used for all of the calculations of spontaneous activity (Figs 1a, b and 2a, b and Supplementary Figs 1 and 4); there were no systematic differences between the interleaved blank data and spontaneous data recorded before any visual stimulation. In some awake protocols, we shortened the interstimulus interval to 300 ms because the evoked responses returned to baseline well before this interval. There was no difference between the longer (1.5 s) and shorter interstimulus interval ($n = 3$) in terms of the magnitude or duration of evoked conductances. For measurements of V_m (Figs 2 and 3 and Supplementary Fig. 3), any spikes were median filtered (7 ms), and the resultant trace was smoothed.

In a subset of experiments, we optimized the orientation of the flashed bars for each individual neuron (Supplementary Fig. 10) by first presenting full-screen drifting gratings (50% contrast, spatial frequency of 0.03 cycles per degree, temporal frequency of 2 Hz and 2 s duration) that varied randomly in orientation (30 degree steps). The preferred orientation was designated as the orientation that evoked the largest number of spikes (or the largest depolarization of V_m), whereas the orthogonal orientation was 90° away from the preferred orientation. Bars that were identical to those in all other experiments (9° width, 100 ms duration, and 1.5 s or 0.3 s interstimulus interval) were then presented randomly across space along the two axes defined by the preferred and orthogonal orientations.

To obtain robust estimates of reversal potential (V_{rev}) and ΔG during the responses, we considered the response onset as the time when ΔG was 5–10% above baseline for 10 ms consecutively. Across the population of neurons, this time point occurred 30–40 ms after stimulus onset, so V_{rev}, ΔG, ΔG_e and ΔG_i were averaged starting 40 ms after stimulus onset, across the population (Fig. 4; $n = 5$ urethane anaesthesia; $n = 8$ awake; both groups were tested identically with vertically oriented bars). Additional recordings of conductances under isoflurane anaesthesia ($n = 1$) and in awake mice that were presented with bars varying in orientation ($n = 4$) were not included in these plots to maintain equivalent group comparisons (but see Supplementary Fig. 9).

The state dependence of visual responses during anaesthesia (Fig. 2) was analysed by sorting trials (within neurons) by the pre-stimulus V_m level. The upper- and lower-most quartiles (that is, the 25% most depolarized and 25% most hyperpolarized non-overlapping trials) were then averaged across the population, and these largely correspond to trials in which the stimuli were delivered in the up state and the down state, respectively[36]. Blank stimuli were also sorted in this manner to estimate the spontaneous state transitions from down to up, and vice versa, in the absence of sensory stimulation.

31. Margrie, T. W., Brecht, M. & Sakmann, B. *In vivo*, low-resistance, whole-cell recordings from neurons in the anaesthetized and awake mammalian brain. *Pflügers Arch.* **444,** 491–498 (2002).
32. Branco, T. & Häusser, M. Synaptic integration gradients in single cortical pyramidal cell dendrites. *Neuron* **69,** 885–892 (2011).
33. Cafaro, J. & Rieke, F. Noise correlations improve response fidelity and stimulus encoding. *Nature* **468,** 964–967 (2010).
34. Poo, C. & Isaacson, J. S. Odor representations in olfactory cortex: 'sparse' coding, global inhibition, and oscillations. *Neuron* **62,** 850–861 (2009).

35. Nowak, L. G., Azouz, R., Sanchez-Vives, M. V., Gray, C. M. & McCormick, D. A. Electrophysiological classes of cat primary visual cortical neurons *in vivo* as revealed by quantitative analyses. *J. Neurophysiol.* **89,** 1541–1566 (2003).

36. Haider, B., Duque, A., Hasenstaub, A. R., Yu, Y. & McCormick, D. A. Enhancement of visual responsiveness by spontaneous local network activity *in vivo*. *J. Neurophysiol.* **97,** 4186–4202 (2007).

LETTER

doi:10.1038/nature11804

Scaling of embryonic patterning based on phase-gradient encoding

Volker M. Lauschke[1]*, Charisios D. Tsiairis[1]*, Paul François[2] & Alexander Aulehla[1]

A fundamental feature of embryonic patterning is the ability to scale and maintain stable proportions despite changes in overall size, for instance during growth[1–6]. A notable example occurs during vertebrate segment formation: after experimental reduction of embryo size, segments form proportionally smaller, and consequently, a normal number of segments is formed[1,7,8]. Despite decades of experimental[1,7] and theoretical work[9–11], the underlying mechanism remains unknown. More recently, ultradian oscillations in gene activity have been linked to the temporal control of segmentation[12]; however, their implication in scaling remains elusive. Here we show that scaling of gene oscillation dynamics underlies segment scaling. To this end, we develop a new experimental model, an *ex vivo* primary cell culture assay that recapitulates mouse mesoderm patterning and segment scaling, in a quasi-monolayer of presomitic mesoderm cells (hereafter termed monolayer PSM or mPSM). Combined with real-time imaging of gene activity, this enabled us to quantify the gradual shift in the oscillation phase and thus determine the resulting phase gradient across the mPSM. Crucially, we show that this phase gradient scales by maintaining a fixed amplitude across mPSM of different lengths. We identify the slope of this phase gradient as a single predictive parameter for segment size, which functions in a size- and temperature-independent manner, revealing a hitherto unrecognized mechanism for scaling. Notably, in contrast to molecular gradients, a phase gradient describes the distribution of a dynamical cellular state. Thus, our phase-gradient scaling findings reveal a new level of dynamic information-processing, and provide evidence for the concept of phase-gradient encoding during embryonic patterning and scaling.

The sequential formation of body segments in vertebrates and arthropod species has been linked to oscillatory gene activity in segment precursor cells[13,14], termed presomitic mesoderm (PSM) in vertebrates. In this context, several signalling pathways, such as Notch[13], Wnt[15] and Fgf[16], have been shown to exhibit oscillatory activity, with periods ranging from ~0.5–2.5 h (ref. 12). Oscillation dynamics can be visualized using *in vivo* real-time imaging methods (Fig. 1a–c and Supplementary Video 1), revealing highly coordinated activity patterns within the PSM. These are kinematic or phase waves, as they result from cell-autonomous oscillatory activity[17] that is phase-shifted between neighbouring cells. Accordingly, physical boundaries do not block their progression (Supplementary Fig. 1 and ref. 17). Although oscillations have been implicated to function within a clock mechanism to control segmentation temporally, the function of the phase shift per se in segment definition has not yet been addressed, in part, owing to the challenge to perturb and analyse phase shifts experimentally at a quantitative level.

We therefore developed an *ex vivo* mouse primary cell culture assay for mesoderm patterning and segment formation, allowing precise quantifications of oscillatory activities in a simplified, two-dimensional system that, in addition, enables us to study segmentation at differing spatial scales.

Using real-time imaging of a dynamic Notch-signalling reporter, LuVeLu[18] (Fig. 1d–g and Supplementary Videos 2 and 3), we found that in this assay, cultured mesoderm cells exhibit robust and long-lasting gene activity oscillations ($n \approx 12$–15 oscillations, Fig. 1d–f). These are equivalent in periodicity to oscillations detected in intact PSM (Fig. 1g). Moreover, we identified highly coordinated, concentric waves of gene activity that sweep across the mPSM in a central–peripheral direction (Fig. 1d). Like their *in vivo* counterparts, these waves are kinematic in nature (Supplementary Fig. 2).

We confirmed that key mesoderm patterning and differentiation events are recapitulated in the *ex vivo* cell culture assay at the molecular level (Fig. 2 and Supplementary Fig. 3). The highest activities of the Wnt- and Fgf-signalling pathways were detected centrally, on the basis of the expression of several direct transcriptional targets, such as *T*, *Msgn1* and *Dusp4*. At the periphery, markers crucial for somite formation and differentiation, such as *Mesp2* and *Uncx4.1* (also known as *Uncx*), become activated (Fig. 2c and Supplementary Fig. 3). Remarkably, we found that at this location in the mPSM, morphological segment boundaries are formed (Fig. 2d, e). Similar to *in vivo*, these boundaries form where gene activity waves halt, coinciding with the expression of *Mesp2* (ref. 19; Supplementary Videos 4–7).

Importantly, we identified that segment scaling occurs after changes in mPSM length. Once the primary cell culture expands to its maximum dimensions after ~20 h of incubation, a progressive decrease in mPSM length is observed. This is due to continuing segmentation which, in the absence of substantial further growth, leads to a constant reduction of undifferentiated mesoderm. Notably, we found that as mPSM length decreases, proportionally smaller segments formed (Fig. 3a). A linear correlation between segment size and mPSM length was maintained ($P < 0.0001$) over the full range of PSM lengths studied (~4-fold changes). In agreement with previous *in vivo* findings in vertebrate embryos[1], segment scaling in our *ex vivo* assay also occurred along the anterior–posterior axis, effectively reducing the width of segments. This demonstrates that segment scaling is recapitulated in our two-dimensional, *ex vivo* cell culture assay, establishing a new experimental model to study the underlying mechanism. To this end, we performed a detailed quantification of oscillations and phase-wave dynamics during the scaling process.

First, we found that in the central mPSM the oscillation period remained constant, irrespective of PSM length (Supplementary Fig. 4). We next analysed the velocity of kinematic waves, a read-out of the phase distribution and therefore of oscillation dynamics in the mPSM (see definition in Supplementary Fig. 5 and below). Surprisingly, we found that the velocities of kinematic waves change linearly with overall mPSM length ($P < 0.0001$), and thus larger samples show proportionally faster kinematic waves than smaller samples (Fig. 3b). This indicated that oscillatory activity is modified to match precisely the spatial context in which it occurs.

To extend this finding, we directly analysed oscillation phases within the mPSM. We used real-time measurements to calculate the phase (φ) independently for every position within the mPSM (Supplementary Fig. 5). In a second step, we determined the spatial phase differences, which allowed us to determine experimentally the phase

[1]Developmental Biology Unit, European Molecular Biology Laboratory, 69117 Heidelberg, Germany. [2]Department of Physics, McGill University, Montréal, Quebec H3A2T8, Canada.
*These authors contributed equally to this work.

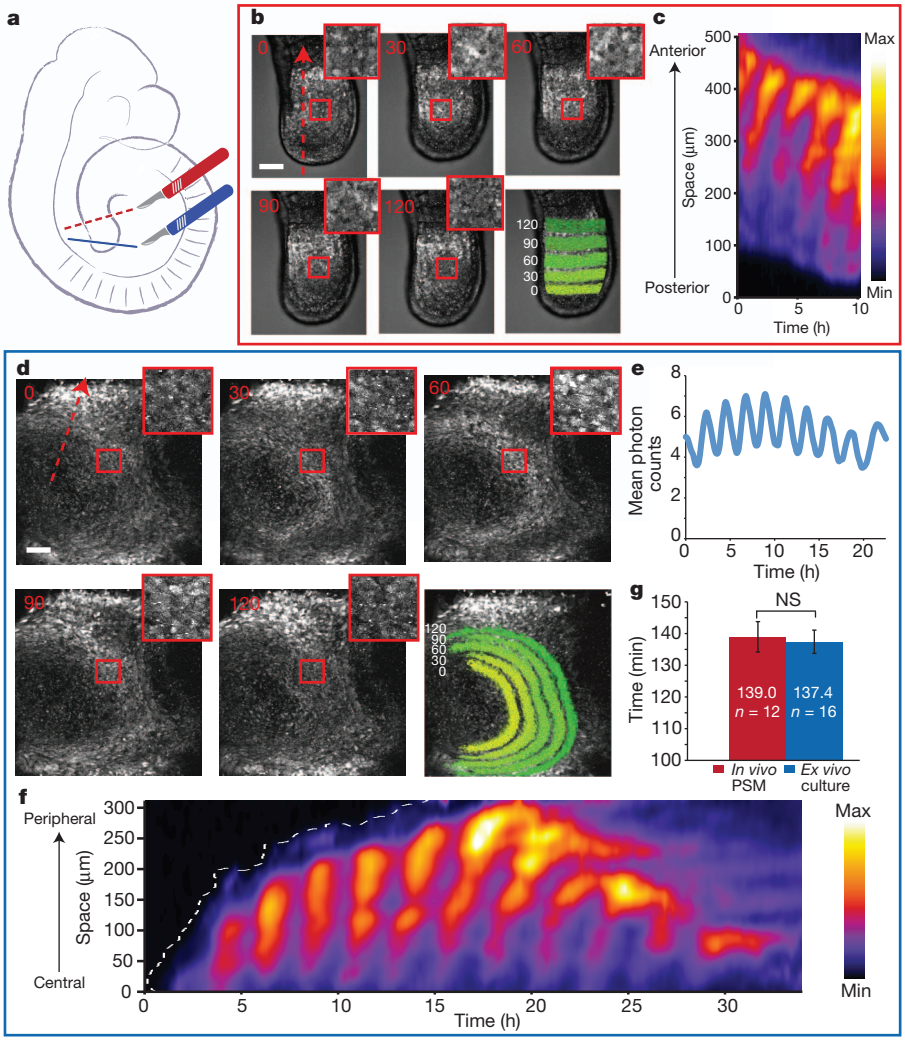


Figure 1 | An *ex vivo* cell culture model for gene activity oscillations. **a**–**c**, Schematic representation of a mouse embryo at embryonic day (E)10.5 (**a**) illustrating the experimental set-up for *in vivo* fluorescence imaging (**b**, **c**) of LuVeLu activity[18]. A posterior embryo fragment including the PSM was used. Dashed red line in **a** shows the level of the cut. **d**–**f**, *Ex vivo* cell culture assay, in which the tail bud mesoderm was isolated (solid blue line in **a**). For the time series of real-time imaging experiments (**b**, **d**), a subregion is magnified (shown at the top right) to exemplify the changes in intensity. The expression patterns at successive time points (times indicated in minutes) are projected into the last thumbnail. **b**, Note periodic activity waves, sweeping from posterior to anterior PSM. **c**, Kymograph along the PSM (arrow in first frame of **b**) showing spatial (*y* axis) and temporal (*x* axis) quantifications (intensity colour-coded), depicting regular oscillations and activity waves from posterior to anterior. **d**, For the *ex vivo* assay, real-time imaging reveals LuVeLu reporter activity waves that progress in a central–peripheral direction. **e**, The intensity within the magnified region in **d** is plotted over time. **f**, Kymograph based on quantification along the arrow depicted in **d**; regular pulses and waves in intensity are seen over the entire course of the *ex vivo* culture. **g**, Quantification of oscillation period in *in vivo* experiments and *ex vivo* cell culture assays shows that there is no significant difference ($P = 0.64$, $n = 12$ for *in vivo* PSM, $n = 16$ for *ex vivo* cultures). Errors bars denote s.d.; NS, not significant. All scale bars, 100 μm.

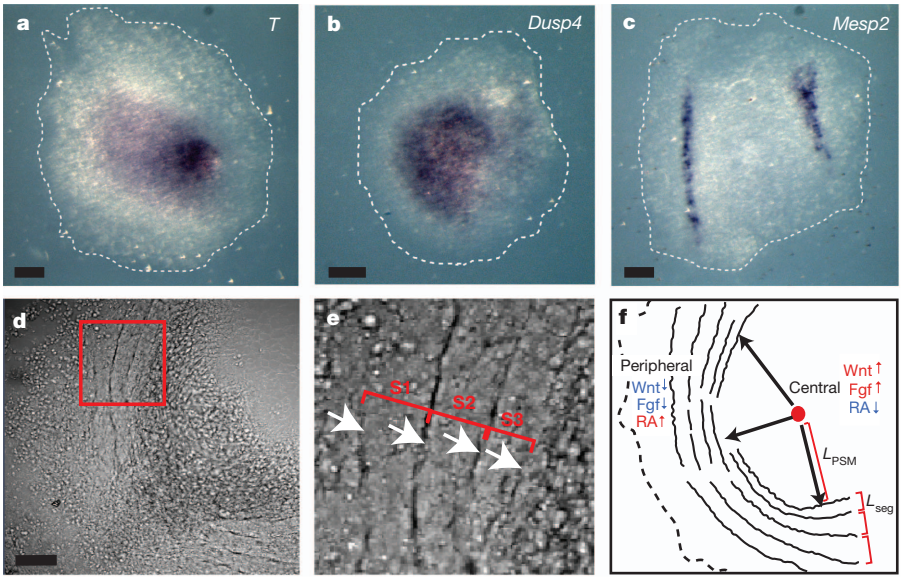


Figure 2 | Molecular and morphological analysis of the *ex vivo* cell culture model using *in situ* hybridization after 18–24 h of culture. **a**, **b**, The Wnt-target gene *T* (**a**) and the Fgf-target gene *Dusp4* (**b**) are expressed centrally. **c**, *Mesp2*, indicative for the onset of mesoderm differentiation, is activated in the periphery of the cell culture assay. **d**, Bright-field image indicating segment formation. **e**, Magnification of region indicated by the red box in **d**, showing sharp boundaries between segments. S1–S3 denote segments in the order of formation. **f**, Scheme illustrating overall reorganization of the embryonic anterior–posterior axis in a central–peripheral direction. Wnt- and Fgf-target genes are upregulated centrally and downregulated peripherally, the inverse is true for the retinoic acid (RA) target gene *Aldh1a2* (Supplementary Fig. 6). L_{PSM} denotes length of PSM; L_{seg} denotes width of one segment. All scale bars, 100 μm.

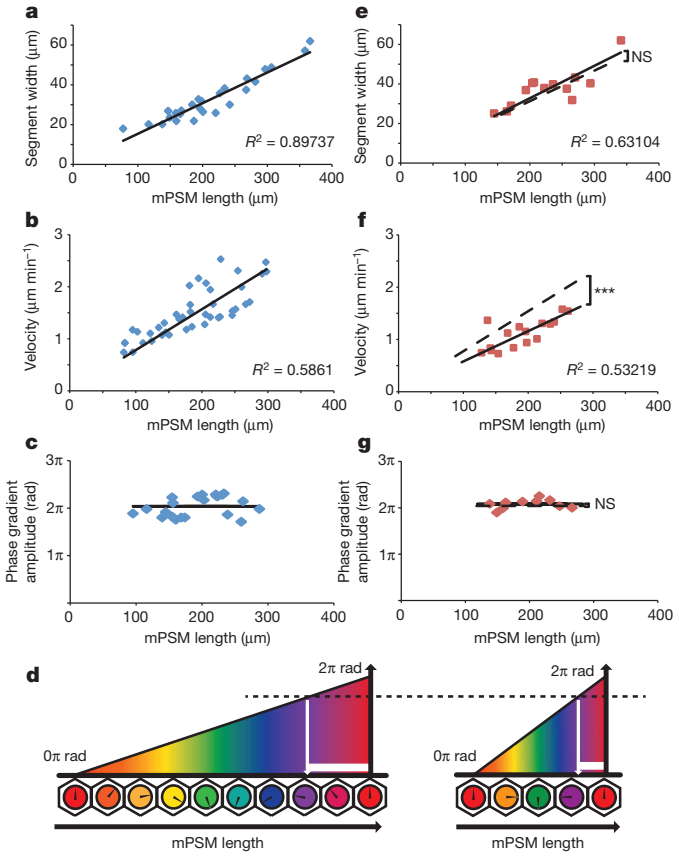


Figure 3 | Quantification of segment sizes and oscillation dynamics reveals scaling behaviour due to a fixed phase gradient amplitude. **a**–**c**, **e**–**g**, Experiments were performed at either 37 °C (**a**–**c**) or 33 °C (**e**–**g**). **a**, Scatter plot of segment widths and length of mPSM from which they form indicates a linear correlation ($P < 0.0001$, $n = 26$ segments; $n = 5$ samples). **b**, The velocity of kinematic waves shows a linear correlation to the length of mPSM in which they occur ($P < 0.0001$, $n = 42$ waves; $n = 7$ samples). As a result, the total time required for a kinematic wave to travel entirely through mPSM of different lengths (time of flight, t_{TOF}) remains constant ($t_{TOF} = 128$ min). **c**, Quantification of the phase-gradient amplitude within mPSM of different lengths: the total amplitude (y axis) is constant ($2.04\pi \pm 0.21\pi$ rad, mean ± s.d., $n = 22$), irrespective of variations in mPSM length (x axis). **d**, Scheme illustrating effect of fixed phase-gradient amplitude in samples of differing size: the total phase span remains constant despite changes in total length, and thus the phase distribution is proportionally adjusted. As a consequence, proportional segments contain the same phase span. **e**, At 33 °C, segment widths show a linear correlation with mPSM length and the regression coefficient is indistinguishable from the correlation found at 37 °C (dashed line; $P = 0.509$, $n = 13$ segments; $n = 4$ samples). **f**, The regression coefficient between kinematic wave velocities and mPSM length is significantly altered after a temperature shift from 37 °C to 33 °C ($P < 0.001$, $n = 16$ waves; $n = 3$ samples). **g**, Even at 33 °C, the phase-gradient amplitude is constant ($2.08\pi \pm 0.1\pi$ rad, $n = 10$), and indistinguishable from that at 37 °C ($P = 0.45$).

gradient in the mPSM (Supplementary Fig. 5). Measuring the phase difference between the most posterior and most anterior oscillating mPSM positions yields the total phase-span present in the mPSM, equivalent to the phase-gradient amplitude. Importantly, we found this amplitude to be constant, close to 2π rad, irrespective of mPSM lengths (Fig. 3c). Having a constant gradient amplitude means that the slope of the phase-gradient ($\partial\varphi/\partial x$, in which x denotes length), the phase difference between mPSM cells, is inversely proportional to mPSM length and segment size. In other words, larger samples show a shallower phase gradient and thus less phase differences between mPSM cells than small samples (Fig. 3d). Such a gradient behaviour can fully account for scaling, as suggested previously[20]. Accordingly, we find that independent of the mPSM length, a newly formed segment spans ~21% of the mPSM phase gradient (Fig. 3a, d).

Our results identify two predictive parameters, the phase-wave velocity (v) and the slope of the phase-gradient ($\partial\varphi/\partial x$), that scale to mPSM length and segment size (Fig. 3b, c). These parameters are interconnected by the following relationship: $v = {\partial\varphi/\partial t}/{\partial\varphi/\partial x}$[21], in which $\partial\varphi/\partial t$ is the oscillation frequency. To dissect the part that each parameter plays during the scaling process, we performed temperature-shift assays. In zebrafish it has been previously shown that a change in temperature influences the rate of segmentation without, however, altering segment sizes[22]. We found a similar result in the mouse mPSM *ex vivo* cell culture model; segmentation proceeded more slowly at 33 °C, yet segment sizes remained unchanged and followed the identical correlation to mPSM length as at 37 °C (Fig. 3e). To test which of the two predictive parameters we identified, v and $\partial\varphi/\partial x$, can account for temperature-invariant segment sizes, we quantified oscillation dynamics in the temperature-shift assay. First, we found that the overall oscillation period (T) was altered by temperature ($T_{37\,°C} = 137.4$ min ± 7.3 s.d., $n = 16$; $T_{33\,°C} = 193.4$ min ± 14.4 s.d., $n = 3$). In addition, we found that although the phase-wave velocity (v) still correlated with mPSM size (Fig. 3f), overall, v was significantly decreased at lower temperature (Fig. 3f). Therefore, phase-wave velocity (v) cannot individually account for temperature-invariant segment size definition.

By contrast, we found that the phase-gradient amplitude remains constant at 2π rad in samples grown at 33 °C (Fig. 3g). This indicates that the phase-gradient slope ($\partial\varphi/\partial x$) is predictive for segment sizes in a temperature-independent manner.

Our findings thus provide strong evidence that a crucial parameter for segment size definition is encoded at the level of phase differences between PSM cells, supporting the general concept that temporal order, that is, a phase gradient, can provide spatial information for embryonic patterning[23].

In principle, the mechanism underlying phase-gradient scaling could rely on sensing global mPSM length, for instance, by integrating signals from opposing signalling gradients[24]. Indeed, such opposing signalling gradients have been previously identified within the PSM[12]. To test this possibility, we analysed our *ex vivo* cell culture assay during the initial culture period (<20 h of culture). Using both morphological and molecular data, we found that at these early time points, only an incomplete mPSM showing uniquely a posterior, undifferentiated mPSM identity, was present (Supplementary Fig. 6). Accordingly, anterior mPSM molecular markers, such as *Mesp2* and *Aldh1a2*, are not yet expressed. Moreover, all cells retain oscillatory activity and segments do not form during the first 20 h of culture. Analysis of the phase-gradient amplitude within the incomplete mPSM showed that it measured clearly less than 2π rad (Supplementary Fig. 7). Importantly, even in the incomplete mPSM and thus in absence of an anterior, opposing gradient, we identified clear evidence for phase-gradient scaling (Fig. 4). Thus, we found that the phase-gradient slope changed exponentially throughout the culture, irrespective of whether a complete mPSM was present or not (Fig. 4b). This suggests the presence of a single underlying mechanism that functions independently of opposing mPSM-signalling gradients.

We then analysed phase-wave velocities within the incomplete mPSM. To this end, we determined the virtual mPSM lengths at which a phase-gradient amplitude of 2π rad, the value that we found to be characteristic for a complete mPSM, is reached (Supplementary Fig. 7 and Fig. 4a). We found that phase-wave velocities are linearly proportional to these projected mPSM lengths (Fig. 4a, inset). Crucially, the correlation is nearly identical to the one found at later time points, when velocities can be compared to de facto, complete mPSM lengths (Fig. 4a, inset).

Such a scaling behaviour is markedly different from other examples of scaling[2,6], as it does not rely on measuring global mPSM size and

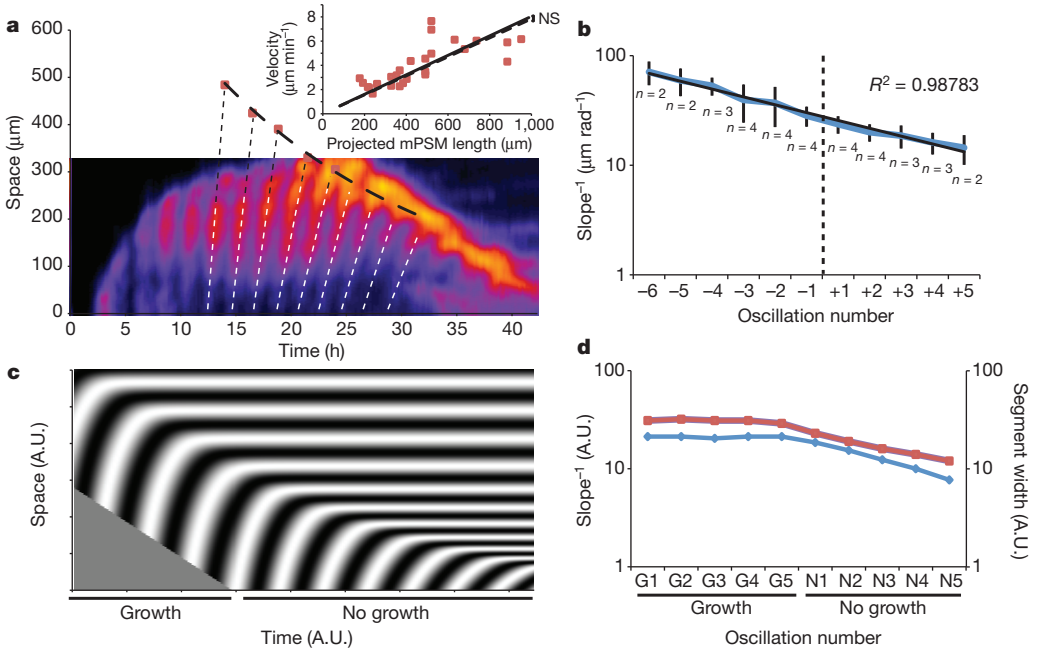


Figure 4 | Phase-gradient scaling does not rely on sensing global size cues. **a**, **b**, Experimental measurements of phase gradients. **c**, **d**, Numerical simulations of the integrative kinematic scaling model (Supplementary Information). **a**, Truncated phase-gradients (waves −5 to −1; see Supplementary Fig. 7) are spatially projected to the point of their completion (red squares); this defines the (projected) mPSM length at which the phase-gradient amplitude measures 2π. Inset shows a scatter plot of early, truncated phase-wave velocities, and projected mPSM lengths indicate a linear correlation ($P < 0.0001$). The regression coefficient is not significantly different ($P = 0.867$, $n = 27$ waves; $n = 6$ samples) from the one calculated at later time points, when actual mPSM lengths are measured (dashed line represents the regression line from Fig. 3b). **b**, Plot of the (inverse) phase-gradient slope (y axis) as a function of oscillations cycles (x axis) shows that the change in slope over time is a uniform function throughout the entire measurements, irrespective of whether a truncated mPSM (waves −6 to −1) or a complete mPSM (waves +1 to +5) is present. An exponential trendline is overlaid in black ($R^2 = 0.99$). Error bars indicate s.d. **c**, Kymograph corresponding to simulation of equation (8) in the Supplementary Information, for two successive regimes (growth and no growth). Levels of grey indicate cosine of phase. The arrested phase serves as a proxy for pattern formation. Note that in this model, the size of the pattern is a fixed fraction of oscillating field size (~PSM). Thus, during PSM growth, patterns are stable (first five waves), once growth ceases and PSM length is reduced, pattern size (~segment size) scales proportionally (last five waves). The growth regime is experimentally observed *in vivo* (Fig. 1b, c), whereas the no-growth regime is reflected experimentally in the *ex vivo* cell culture assay (**a**). A.U., arbitrary units. **d**, On the basis of numerical simulations, the slope of the phase gradient (plotted as the inverse, blue diamonds) was calculated for each consecutive oscillation cycle (x axis). Note the (inverse) correlation of the phase-gradient slope to corresponding segment size (red squares).

operates even when the total field to be scaled is not yet completely formed.

Furthermore, these findings reflect the ability of mPSM to retain a developmental memory after isolation, as has been previously found in classical experimental embryology work (reviewed in ref. 25), and as has been suggested in theoretical models[20]. Accordingly, segments form after an initial delay of ~20 h (Fig. 4) as the posterior mPSM fragments retain an undifferentiated identity. At the same time, however, our approach reveals a marked plasticity within the posterior mPSM that enables considerable changes in oscillatory dynamics and ultimately, segment scaling to occur.

The precise molecular circuits leading to segment scaling are as yet unknown. Previously, signalling gradients were functionally implicated in controlling PSM differentiation[12], however, experimental evidence for a role in directly controlling the oscillation dynamics is still lacking. We confirmed that Wnt and Fgf gradients are present throughout the *ex vivo* culture (Fig. 2 and Supplementary Fig. 8). However, in the absence of quantitative and, importantly, dynamic data, their precise function in this scaling process and the interaction with complex phase-gradient dynamics awaits further investigation.

Here we used modelling to address the principles of segment scaling. We mathematically formalized our results and asked whether these are sufficient to recover scaling behaviour *in silico*. This model (see full derivation in Supplementary Information) is based on the following key findings. First, we identified that oscillation dynamics pivot around a fixed phase-gradient amplitude of ~2π rad. The observed maintenance of a fixed phase shift $\Delta\varphi_*$ between the posterior PSM and the anterior differentiation front, irrespective of PSM length, strongly argues for a tight interdependence between segmentation oscillator and differentiation front definition. This is in contrast to common models for segmentation, such as the clock and wave-front model[10], in which oscillations and differentiation front are two interacting, but essentially independent entities. Second, our quantifications indicate that the phase gradient evolves nonlinearly over time, with an exponential function describing the system accurately (Fig. 4b). This means that oscillations are slowing down exponentially throughout the entire culture period (with the exception of the most posterior mPSM location; Supplementary Fig. 4). These findings were formalized using a kinematic phase equation[26]. Taking further into account posterior growth and coupling between cells in the PSM (Supplementary Information), the complete numerical simulation closely recapitulates the observed experimental data (Fig. 4c, d). This shows that this minimal set of principles can fully account for segment scaling. Thus, we propose that scaling is a direct consequence of nonlinear phase-gradient dynamics combined with a fixed gradient amplitude.

Finally, our identification that the phase gradient is the single-known predictive parameter for both segment scaling and temperature compensation has important conceptual implications. Phase differences within the PSM are a defining feature of oscillations during the segmentation process[13,14], and the mechanism of how these differences are set up is being investigated[27–29]. However, current models of segmentation[10,30] generally do not assign any clear functional relevance to these phase differences per se. This work prompts a change in perspective as our experimental findings support the concept that

spatial patterning is based on information encoded at the level of oscillation phase differences between mPSM cells.

In conclusion, we present a powerful *ex vivo* culture assay for mesoderm segmentation that enabled the identification of a conceptually distinct segment-scaling mechanism based on phase-gradient encoding.

METHODS SUMMARY

All parameter measurements were done on data from real-time imaging of the *LuVeLu* transgenic reporter line[18], which expresses a destabilized version of Venus (yellow fluorescent protein) under the lunatic fringe reporter. For the *ex vivo* cell culture assay, tail bud mesoderm was removed from mouse embryos at embryonic day 10.5 and cultured in fibronectin-coated chamber slides under controlled environmental conditions on an inverted microscope (Zeiss LSM 780 NLO). *In situ* messenger RNA hybridizations and immunohistochemistry were performed directly in chamber slides. Image processing, analysis and parameter measurements were done using software packages ZEN (Zeiss), Fiji and MATLAB.

Full Methods and any associated references are available in the online version of the paper.

Received 7 May; accepted 19 November 2012.
Published online 19 December 2012.

1. Cooke, J. Control of somite number during morphogenesis of a vertebrate, *Xenopus laevis*. *Nature* **254,** 196–199 (1975).
2. Gregor, T., Bialek, W., de Ruyter van Steveninck, R. R., Tank, D. W. & Wieschaus, E. F. Diffusion and scaling during early embryonic pattern formation. *Proc. Natl Acad. Sci. USA* **102,** 18403–18407 (2005).
3. Hamaratoglu, F., de Lachapelle, A. M., Pyrowolakis, G., Bergmann, S. & Affolter, M. Dpp signaling activity requires Pentagone to scale with tissue size in the growing *Drosophila* wing imaginal disc. *PLoS Biol.* **9,** e1001182 (2011).
4. Ben-Zvi, D., Shilo, B. Z., Fainsod, A. & Barkai, N. Scaling of the BMP activation gradient in *Xenopus* embryos. *Nature* **453,** 1205–1211 (2008).
5. Restrepo, S. & Basler, K. Morphogen gradients: expand and repress. *Curr. Biol.* **21,** R815–R817 (2011).
6. Wartlick, O. *et al.* Dynamics of Dpp signaling and proliferation control. *Science* **331,** 1154–1159 (2011).
7. Tam, P. P. The control of somitogenesis in mouse embryos. *J. Embryol. Exp. Morphol.* **65** (suppl.), 103–128 (1981).
8. Brown, D. *et al.* Loss of Aif function causes cell death in the mouse embryo, but the temporal progression of patterning is normal. *Proc. Natl Acad. Sci. USA* **103,** 9918–9923 (2006).
9. Wolpert, L. Positional information and the spatial pattern of cellular differentiation. *J. Theor. Biol.* **25,** 1–47 (1969).
10. Cooke, J. & Zeeman, E. C. A clock and wavefront model for control of the number of repeated structures during animal morphogenesis. *J. Theor. Biol.* **58,** 455–476 (1976).
11. Meinhardt, H. in *Somites in Developing Embryos* (eds Bellairs, R., Ede, D. A. & Lash, J. W.) 179–191 (Plenum, 1986).
12. Dequéant, M.-L. & Pourquié, O. Segmental patterning of the vertebrate embryonic axis. *Nature Rev. Genet.* **9,** 370–382 (2008).
13. Palmeirim, I., Henrique, D., Ish-Horowicz, D. & Pourquié, O. Avian hairy gene expression identifies a molecular clock linked to vertebrate segmentation and somitogenesis. *Cell* **91,** 639–648 (1997).
14. Sarrazin, A. F., Peel, A. D. & Averof, M. A segmentation clock with two-segment periodicity in insects. *Science* **336,** 338–341 (2012).
15. Aulehla, A. *et al.* Wnt3a plays a major role in the segmentation clock controlling somitogenesis. *Dev. Cell* **4,** 395–406 (2003).
16. Niwa, Y. *et al.* The initiation and propagation of Hes7 oscillation are cooperatively regulated by Fgf and Notch signaling in the somite segmentation clock. *Dev. Cell* **13,** 298–304 (2007).
17. Masamizu, Y. *et al.* Real-time imaging of the somite segmentation clock: revelation of unstable oscillators in the individual presomitic mesoderm cells. *Proc. Natl Acad. Sci. USA* **103,** 1313–1318 (2006).
18. Aulehla, A. *et al.* A β-catenin gradient links the clock and wavefront systems in mouse embryo segmentation. *Nature Cell Biol.* **10,** 186–193 (2008).
19. Morimoto, M., Takahashi, Y., Endo, M. & Saga, Y. The Mesp2 transcription factor establishes segmental borders by suppressing Notch activity. *Nature* **435,** 354–359 (2005).
20. Meinhardt, H. *Models of Biological Pattern Formation* (Academic, 1982).
21. Ross, J., Müller, S. C. & Vidal, C. Chemical waves. *Science* **240,** 460–465 (1988).
22. Schröter, C. *et al.* Dynamics of zebrafish somitogenesis. *Dev. Dyn.* **237,** 545–553 (2008).
23. Goodwin, B. C. & Cohen, M. H. A phase-shift model for the spatial and temporal organization of developing systems. *J. Theor. Biol.* **25,** 49–107 (1969).
24. McHale, P., Rappel, W. J. & Levine, H. Embryonic pattern scaling achieved by oppositely directed morphogen gradients. *Phys. Biol.* **3,** 107–120 (2006).
25. Keynes, R. J. & Stern, C. D. Mechanisms of vertebrate segmentation. *Development* **103,** 413–429 (1988).
26. Kopell, N. & Howard, L. N. Horizontal bands in the Belousov reaction. *Science* **180,** 1171–1173 (1973).
27. Horikawa, K., Ishimatsu, K., Yoshimoto, E., Kondo, S. & Takeda, H. Noise-resistant and synchronized oscillation of the segmentation clock. *Nature* **441,** 719–723 (2006).
28. Riedel-Kruse, I. H., Muller, C. & Oates, A. C. Synchrony dynamics during initiation, failure, and rescue of the segmentation clock. *Science* **317,** 1911–1915 (2007).
29. Morelli, L. G. *et al.* Delayed coupling theory of vertebrate segmentation. *HFSP J.* **3,** 55–66 (2009).
30. Roellig, D., Morelli, L. G., Ares, S., Julicher, F. & Oates, A. C. SnapShot: the segmentation clock. *Cell* **145,** 800–800 (2011).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank members of the Aulehla laboratory for discussion and comments on the manuscript. We thank M. Snaebjornsson for providing Supplementary Video 1. We thank F. Peri, D. Gilmour, T. Hiiragi and F. Spitz for comments on the manuscript and P. Riedinger for artwork. This work was supported by EMBL Imaging and Laboratory animal resource core facilities. The *Mesp2-GFP* line was provided by Y. Saga. P.F. was supported by Natural Science and Engineering Research Council of Canada (NSERC), Discovery Grant program RGPIN 401950-11, Regroupement Québécois pour les matériaux de pointe (RQMP) and McGill University.

Author Contributions V.M.L. developed the *ex vivo* culture system, performed the experiments and analysed the data; C.D.T. performed the experiments, analysed the data and developed the oscillation-phase quantification; P.F. developed and wrote the mathematical model; A.A. designed and supervised the project and wrote the manuscript. All authors discussed and contributed to the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to A.A. (aulehla@embl.de).

METHODS

***Ex vivo* cell culture assay for mesoderm patterning.** Chamber slides (Lab-Tek) were coated with 50 μg ml^{-1} fibronectin (Sigma) in 100 mM NaCl for 4 h at room temperature or overnight at 4 °C. After coating, the dishes were air-dried for 5 min and washed with embryo culture medium (DMEM-F12, Cell Culture Technologies) plus 0.5 mM glucose, 2 mM glutamine and 1% BSA) for 20 min with slight shaking. E10.5 embryos of *LuVeLu*^{het} × *Cd1* crosses were dissected in HEPES-buffered embryo culture medium (culture medium plus 10 mM HEPES). To isolate tailbud mesoderm, a transversal cut at or behind the posterior neuropore was done. Tailbuds were placed with this transverse cut facing down on the fibronectin-coated dishes. Routine culture conditions were 37 °C, 5% CO_2 and ambient O_2 levels, or as previously described[18].

Two-photon microscopy. Imaging was performed with a Zeiss LSM780 laser-scanning microscope. Samples were excited using a Ti:Sapphire Laser (Chameleon-Ultra, Coherent) at a wavelength of 960 nm through a ×20 plan apo objective (numerical aperture 0.8). A *Z*-stack of either 6–8 planes (for *in vivo* imaging) or 3–4 planes (for the *ex vivo* assay) at a 6–8 μm distance was scanned every 10 min; up to 16 samples were recorded during an experiment using a motorized stage.

Further mouse strains and animal work. The *Mesp2-GFP* line[31] was obtained by the RIKEN BRC through the National Bio-Resource Project of the MEXT, Japan. Both the *LuVeLu* and the *Mesp2-GFP* line were maintained on a *Cd1* background. All animal experiments were conducted under veterinarian supervision and rules of the European Molecular Biology Laboratory, following the guidelines of the European Commission, Directive 2010/63/EU and AVMA Guidelines 2007.

***In situ* hybridization on primary culture PSM tissue.** *In situ* hybridization and probe generation was performed as described previously[18], with minor modifications when applied to the primary *ex vivo* culture samples (that is, reduced incubation and washing times). The entire hybridization and colour reactions were performed in chamber slides. Probes were used as described previously[18] or amplified using the following primers: *Meox1* forward 5′-TGAGATTGCAGTCA ACCTGG-3′, reverse 5′-TTGGAGAACACAAGACGCTG-3′; *Pitx2* forward 5′-TCAGAGTATGTTTTCCCCGC-3′, reverse 5′-CGCAAGCGAAAAATCCTAAC-3′; *Cldn6* forward 5′-CAGAGCCCTCTGTGTTGTCA-3′, reverse 5′-AGAGGTGG AGCTTGGACTCA-3′; *Sox2* forward 5′-ACCAGCTCGCAGACCTACAT-3′, reverse 5′-ACGAAAACGGTCTTGCCAGT-3′.

Immunofluorescence on primary culture PSM tissue. *Ex vivo* cell cultures were washed with PBS and fixed on ice for 1 min in 50% methanol, 50% dimethylsulphoxide (DMSO). Afterwards, cells were treated with 15% H_2O_2, 50 mM NH_4Cl in PBS for 15 min on ice. Cells were washed three times with 1% Triton X-100 in PBS and blocked for 10 min with PBS plus 1% Triton X-100 and 10% FCS. Primary antibody incubations were carried out at 4 °C overnight. Phosphorylated ERK1/2 was detected with a 1:150 dilution of rabbit-anti-phospho-p42/44 MAPK (ERK1/2) (Cell Signaling, D13.14.4), β-catenin was detected using a 1:500 dilution of mouse-anti-β-catenin (BD Transduction, 610153). After incubation, the cells were washed three times with PBS plus 1% Triton X-100 and 10% FCS, and subsequently incubated with secondary antibody (1:250 of Fab_2-goat-anti-rabbit-Alexa488; Cell Signaling, 4412S; or 1:500 of Fab_2-goat-anti-mouse-Alexa555; Cell Signaling, 4409S) at 4 °C overnight. After secondary antibody incubation, cells were washed twice with PBS and 1% Triton X-100 plus 10% FCS, and incubated with 1:1,000 4′,6-diamidino-2-phenylindole (DAPI) in PBS plus 1% Triton X-100 for 1 h at room temperature. Once nuclear staining was judged to be sufficient, cells were imaged on an LSM780 microscope (Zeiss).

For temperature-shift experiments, samples were grown at 33 °C; only samples in which clear segment formation occurred were used for subsequent analysis.

Image processing. All data were recorded in 12bit, 512 × 512 pixels, 1.38 μm pixel^{-1}, all *z*-planes were considered for maximum intensity projections done in the ZEN-software. Fiji[32]was used for further image processing and the generation of kymographs. To this end, maximum projection time series were blurred using a Gaussian filter (7 pixel sigma radius). Kymographs were generated along a line from the oscillation centre to the periphery on the blurred videos, perpendicular to the wavefront. For further analysis with MATLAB, the kymographs were blurred again (Gaussian, 7 pixel sigma radius) and saved as text image.

The trend of the kymograph intensity was removed and the intensity was smoothened using a moving average function. The Hilbert transform[33] was calculated along the time coordinate for every space point. From the Hilbert transform the instantaneous phase was extracted and plotted for every location and for all time points (Supplementary Fig. 5).

Images of *in situ* mRNA hybridizations were taken using a Leica MZ16F stereomiscroscope and a Leica DFC420C digital camera. Brightness and contrast were nonlinearly adjusted uniformly to the entire image.

Parameter measurements. Velocities of kinematic waves were measured on the basis of phase kymographs. For each wave, a line was manually placed over the calculated phase with a value of 2π. The resulting speed was given by tan(ϑ) × time unit per pixel/space unit per pixel.

Segment widths were measured at defined time points using ZEN software based on identification of segment boundaries in the transmitted light image; three independent measurements were taken and averaged.

mPSM length was measured as the distance between the point of origin of oscillations (or, in *LuVeLu*-negative explants, corresponding morphological landmarks), and the positions at which these oscillations halt.

To calculate the phase gradient and the total phase shift (or amplitude) in the mPSM ($\Delta\varphi_o$), phase kymographs were used to unwrap phases along the space dimension for every time value. Then, $\Delta\varphi_o = \varphi_{max} - \varphi_{min}$ for time points at which wave troughs reach the end of the mPSM. The average $\Delta\varphi_o$ for 37 °C was calculated from 22 waves from seven samples, and for 33 °C from 10 waves from three samples.

From phase kymographs the slopes $\Delta\varphi/\Delta x$ (which in approximation corresponds to the partial derivative $\partial\phi/\partial x$) were calculated. Phase values along space were plotted for time points at which each wave's trough reaches the anterior mPSM end, as defined in phase kymographs. Correlation of phase (φ) as a function of length (x) was close to linear for the samples grown at 37 °C, and the slope ($\Delta\varphi/\Delta x$) was calculated by linear fitting.

The time of flight (t_{TOF}) was determined using the regression line in Fig. 3b, t_{TOF} being the inverse of the slope. Periods for selected regions of interest were calculated using Lomb–Scargle algorithm[34] applied on the changes of intensity over time.

Statistics. Regression lines were fitted in Excel (Microsoft). For graphical representation only, lines were forced through zero. Goodness of fit (R^2), probability of linear correlation (F test) and similarity of regression coefficients (analysis of covariance; ANCOVA) were calculated with Prism 6 (GraphPad software). Comparison of means of two samples (unpaired Student's t-test) was performed with Excel (Microsoft).

31. Morimoto, M., Kiso, M., Sasaki, N. & Saga, Y. Cooperative Mesp activity is required for normal somitogenesis along the anterior–posterior axis. *Dev. Biol.* **300,** 687–698 (2006).
32. Schindelin, J. *et al.* Fiji: an open-source platform for biological-image analysis. *Nature Methods* **9,** 676–682 (2012).
33. Pikovsky, A., Rosenblum, M. & Kurths, J. *Synchronization: a Universal Concept in Nonlinear Sciences* (Cambridge Univ., 2001).
34. Glynn, E. F., Chen, J. & Mushegian, A. R. Detecting periodic patterns in unevenly spaced gene expression time series using Lomb–Scargle periodograms. *Bioinformatics* **22,** 310–316 (2006).

LETTER

doi:10.1038/nature11693

Restriction of intestinal stem cell expansion and the regenerative response by YAP

Evan R. Barry[1,2,3], Teppei Morikawa[4], Brian L. Butler[1,2], Kriti Shrestha[1], Rosemarie de la Rosa[1], Kelley S. Yan[5], Charles S. Fuchs[4,6], Scott T. Magness[7], Ron Smits[8], Shuji Ogino[4,9], Calvin J. Kuo[5] & Fernando D. Camargo[1,2,3]

A remarkable feature of regenerative processes is their ability to halt proliferation once an organ's structure has been restored. The Wnt signalling pathway is the major driving force for homeostatic self-renewal and regeneration in the mammalian intestine. However, the mechanisms that counterbalance Wnt-driven proliferation are poorly understood. Here we demonstrate in mice and humans that yes-associated protein 1 (YAP; also known as YAP1)—a protein known for its powerful growth-inducing and oncogenic properties[1,2]—has an unexpected growth-suppressive function, restricting Wnt signals during intestinal regeneration. Transgenic expression of YAP reduces Wnt target gene expression and results in the rapid loss of intestinal crypts. In addition, loss of YAP results in Wnt hypersensitivity during regeneration, leading to hyperplasia, expansion of intestinal stem cells and niche cells, and formation of ectopic crypts and microadenomas. We find that cytoplasmic YAP restricts elevated Wnt signalling independently of the AXIN–APC–GSK-3β complex partly by limiting the activity of dishevelled (DVL). DVL signals in the nucleus of intestinal stem cells, and its forced expression leads to enhanced Wnt signalling in crypts. YAP dampens Wnt signals by restricting DVL nuclear translocation during regenerative growth. Finally, we provide evidence that YAP is silenced in a subset of highly aggressive and undifferentiated human colorectal carcinomas, and that its expression can restrict the growth of colorectal carcinoma xenografts. Collectively, our work describes a novel mechanistic paradigm for how proliferative signals are counterbalanced in regenerating tissues. Additionally, our findings have important implications for the targeting of YAP in human malignancies.

YAP is a critical component of the Hippo signalling pathway, which controls organ size in mammals[1,2]. Through a kinase cascade, the pathway targets YAP for phosphorylation, preventing its translocation to the nucleus, where it functions as a transcriptional co-activator. Current dogma suggests that restriction of the transcriptional activity of YAP is the principal mechanism of growth and tumour suppression by the Hippo pathway[2]. Indeed, nuclear YAP is a powerful driver of organ growth, progenitor proliferation and tumour growth[1–4]. We previously assessed YAP function in the mammalian intestine by using a mouse model that resulted in ubiquitous postnatal expression of an inducible YAP(S127A) mutant[3]. This mutant protein is thought to have enhanced nuclear localization given that it escapes inactivation by the Hippo kinases LATS1 and LATS2 (ref. 3). As YAP might activate paracrine signals[5], we sought to bypass non-cell-autonomous effects by specifically expressing YAP in the intestinal epithelium using the Villin-rtTA driver[6]. YAP protein in the intestine of transgenic mice was not restricted to the nucleus, suggesting that S127 is not the major determinant of YAP subcellular localization in this tissue (Supplementary Fig. 1a). Five-to-seven days after doxycycline (Dox) administration, transgenic mice became moribund and were killed. Surprisingly, histological evaluation of the small intestine and colon of transgenic mice revealed a progressive degenerative phenotype associated with the rapid loss of proliferating crypts (Fig. 1a and Supplementary Fig. 1b, c).

Crypt-loss phenotypes are typically associated with reduced Wnt signalling[7]. Indeed, degeneration was accompanied by repression of the Wnt target gene *Cd44* and loss of cells with nuclear β-catenin (Fig. 1b and Supplementary Fig. 1e–g). Paneth cells are a mature intestinal lineage that require high levels of Wnt signalling for their proper differentiation and localization, and function as a critical component of the intestinal stem cell (ISC) niche[8,9]. In YAP transgenic

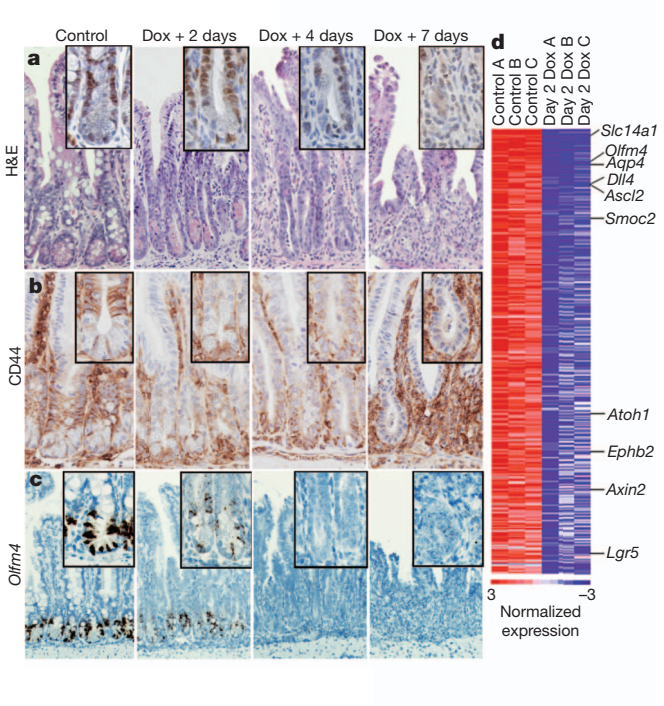


Figure 1 | YAP overabundance inhibits Wnt-mediated intestinal regeneration. **a**, Haematoxylin and eosin (H&E) staining of Dox-induced YAP(S127A) small intestine at 2, 4 and 7 days. Inset shows Ki67 stain representative of crypt proliferation. **b**, **c**, Wnt pathway activity and ISC presence at 2, 4 and 7 days after Dox induction represented by CD44 (**b**) and *Olfm4* (*in situ*) (**c**). **d**, Heatmap of crypts isolated from control ($n = 3$) and day 2 Dox-treated ($n = 3$) mice showing 540 downregulated genes in rank order. Labelled genes are examples known to be involved in Wnt signalling and the ISC niche. Original magnifications in each panel are ×20 (**a**, **b**), ×10 (**c**).

[1]Stem Cell Program and Department of Hematology/Oncology, Children's Hospital, Boston, Massachusetts 02115, USA. [2]Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, Massachusetts 02138, USA. [3]Harvard Stem Cell Institute, Cambridge, Massachusetts 02138, USA. [4]Department of Medical Oncology, Dana-Farber Cancer Institute and Harvard Medical School, Boston, Massachusetts 02215, USA. [5]Department of Medicine, Hematology Division, Stanford University, Stanford, California 94305, USA. [6]Channing Laboratory, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts 02115, USA. [7]Departments of Medicine and Biomedical Engineering, The University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA. [8]Department of Gastroenterology and Hepatology, Erasmus MC, 3000 CA Rotterdam, The Netherlands. [9]Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts 02115, USA.

mice, Paneth cells become mislocalized and eventually disappear (Supplementary Fig. 1d). To determine if YAP expression was reducing ISC numbers, we performed *in situ* hybridization (ISH) for *Olfm4*, which marks crypt base columnar (CBCs) stem cells[10]. *Olfm4*$^{+}$ CBCs were markedly reduced 2 days after induction and were essentially absent by day 4 (Fig. 1c). Consistent with these data, microarray analysis on isolated crypts showed that many robustly downregulated transcripts were known ISC signature genes (that is, *Olfm4*, *Ascl2*, *Smoc2* and *Lgr5*) (Fig. 1d and Supplementary Fig. 2a, b). Gene set enrichment analysis (GSEA) demonstrated significant downregulation of both intestinal β-catenin targets and a recently described ISC gene signature[11,12] (Supplementary Fig. 1c and Supplementary Table 1). Inhibition of growth, loss of Paneth cells and suppression of Wnt/ISC signature genes were confirmed in organoid cultures derived from transgenic mice (Supplementary Fig. 2d, e)[9]. Together, these results demonstrate that, in contrast to other tissues where it promotes growth, epithelial-specific expression of YAP suppresses intestinal renewal, which occurs through inhibition of the Wnt signalling pathway.

These observations prompted us to evaluate more carefully the phenotype of gut-specific YAP mutant mice. Consistent with a previous report[13], developmental or acute YAP loss produced no major abnormalities during normal intestinal homeostasis (Fig. 2b and Supplementary Fig. 3a). However, after injury by whole-body irradiation (9 and 11 Gy), *Villin-Cre Yap*$^{fl/fl}$ (conditional knockout (cKO)) mice showed a phenotype of crypt hyperplasia and overgrowth throughout the small intestine and colon (Fig. 2a and Supplementary Fig. 3c, e). This observation contrasts to that of the impaired repair observed in a dextran sulphate sodium salt (DSS)-mediated colitis model[13] (Supplementary Fig. 3b). cKO crypts were hyperproliferative and showed upregulation of the Wnt target genes *Cd44* and *Sox9* as well as mislocalized and increased numbers of Paneth cells (Fig. 2a and Supplementary Fig. 3f). Apoptosis was not altered in cKO mice (Supplementary Fig. 3d).

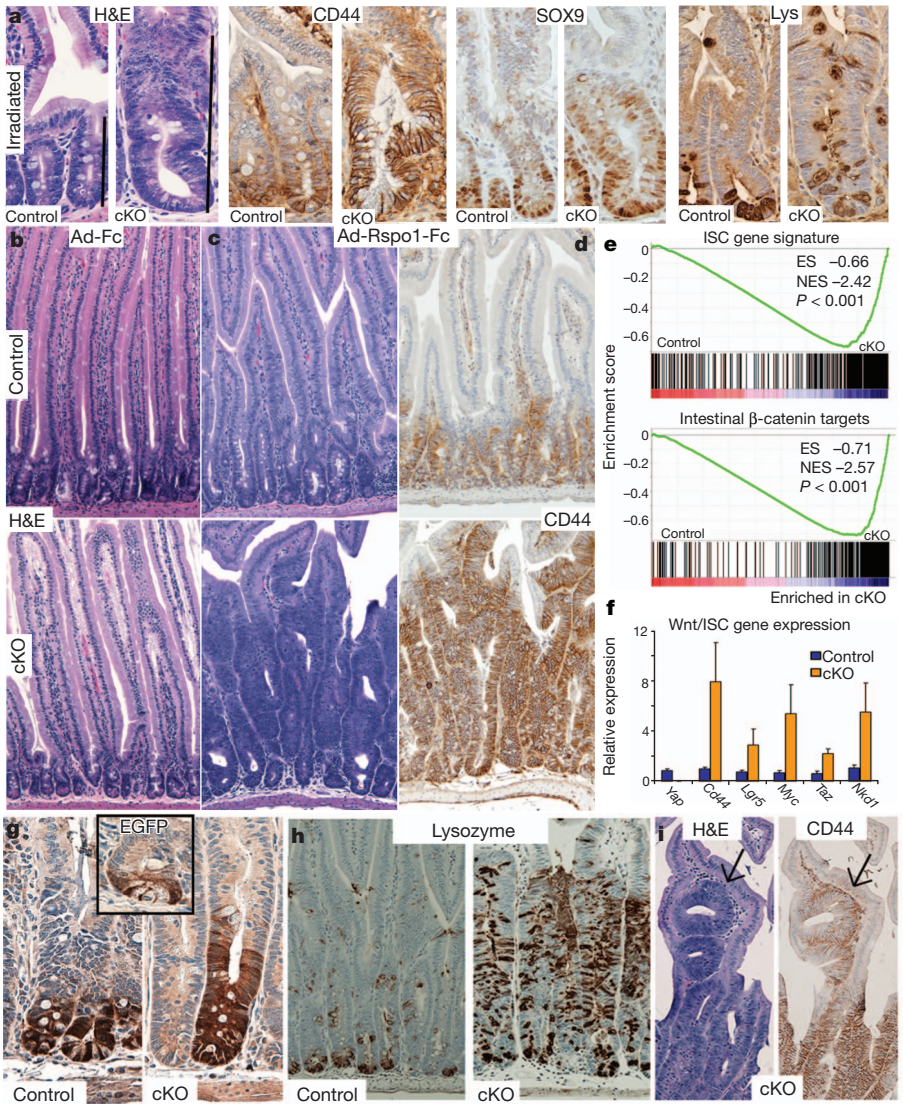


Figure 2 | Loss of YAP leads to hyperactive Wnt signalling and expansion of the stem cell niche after injury or stimulation with Rspo1. **a**, H&E, CD44, SOX9 and lysozyme (Lys) staining of small intestinal crypts in control and cKO mice 1 week after irradiation. **b**, H&E staining of inert-virus-treated control and cKO small intestine. **c**, **d**, H&E staining (**c**) and CD44 immunohistochemistry (**d**) in control and cKO mice 1 week after administration of adenovirus expressing Rspo1. **e**, GSEA of the ISC gene signature and intestine-specific β-catenin target gene sets. Black bars represent individual genes in rank order. ES, enrichment score, NES, normalized enrichment score. **f**, Quantitative polymerase chain reaction (qPCR) validating several upregulated Wnt/ISC markers. **g**, **h**, Immunohistochemistry on control and cKO small intestine for EGFP (expressed from the Lgr5 locus, inset is an untreated mouse) (**g**) and lysozyme marking Paneth cells (**h**). **i**, Ectopic crypt formation in cKO mice treated with Rspo1 stained for H&E or the Wnt target CD44. Original magnifications in each panel are ×20 (**a**), ×10 (**b**–**d**), ×20 (**g**) and ×10 (**h**, **i**). Graphed data represent the mean and standard error of mean of three individual mice per genotype.

Considering that intestinal regeneration after irradiation is characterized by a state of Wnt hyperactivity[14,15], our data suggest a role for YAP in restricting elevated Wnt signalling *in vivo*.

To test directly if loss of YAP results in Wnt hypersensitivity, we injected adenovirus expressing R-spondin1 (Ad-Rspo1; R-spondin1 also known as Rspo1) into cKO and control mice. Rspo1 is a potent secreted Wnt agonist that induces crypt-cell proliferation[16,17]. Seven days after Ad-Rspo1 injection, 80% of cKO mice ($n = 8$) became moribund and were killed, whereas Ad-Fc-injected control mice ($n = 8$) appeared normal. The intestines and colon of cKO mice showed a massively hyperplastic phenotype, so that by day 7 most mature intestinal lineages in cKO epithelia were replaced by highly proliferative crypt-like tissue (Fig. 2b, c and Supplementary Fig. 4a–i). Hyperplasia was accompanied by upregulation of Wnt targets CD44, SOX9 and EPHB3, in addition to global upregulation of the intestinal β-catenin target signature (Fig. 2d–f and Supplementary Fig. 4j). To determine if YAP loss resulted in expansion of cells with the features of ISCs, we generated cKO:*Lgr5-EGFP* mice. LGR5 is normally expressed in the CBCs at the very bottom of the crypt (Fig. 2g, inset); however, after Rspo1 administration in control mice, the population of $Lgr5^+$ ISCs expands (Fig. 2g). This expansion is much more marked in the cKO intestine, where the $Lgr5^+$ domain is 3–4 times larger. ISC expansion was confirmed by ISH for *Olfm4* (Supplementary Fig. 5a), and by GSEA (Fig. 2e and Supplementary Table 1). cKO mice also showed a prominent increase in Paneth cell numbers (Fig. 2h and Supplementary Fig. 5b, c, d). Paneth cells and $LGR5^+$ cells were evident high in the crypt/villus axis (Fig. 2g and Supplementary Fig. 5c, d). We also observed ectopic epithelial foci resembling crypts in cKO mice, which had formed within villi and stained positive for CD44, lysozyme and proliferative markers (Fig. 2i and Supplementary Fig. 5e–g). Some of these progressed into structures reminiscent of microadenomas, which are pre-cancerous lesions of outpocketing pouches of crypt-like tissue that grow into normal villi (Supplementary Fig. 5g). These were never observed in control Rspo1-treated animals. Together, these results demonstrate that YAP restricts the expansion of ISCs as well as critical components of the stem cell niche.

To gain insight into the mechanism behind our observations, we revisited the expression pattern of YAP in the intestine. It has previously been shown that YAP is primarily localized to the crypt base, and is absent from villi[3,13]. Using a new polyclonal antibody, we find that YAP is enriched in the nucleus of CBCs and lower crypt cells (Supplementary Fig. 6a–c) but was also expressed in upper transit amplifying (TA) cells and throughout the villi, where it is predominantly cytoplasmic (Supplementary Fig. 6a). This was confirmed by cellular fractionation experiments (Supplementary Fig. 6d). YAP localization correlated with the staining pattern of CD44, demonstrating that nuclear YAP is present in zones of active Wnt signalling, but is mostly cytoplasmic where Wnt is restricted (Supplementary Fig. 6e). These data indicate that cytoplasmic YAP might be functionally responsible for terminating Wnt signalling and allowing progress from a proliferative progenitor/stem cell compartment to a post-mitotic, differentiated fate. To rule out the possibility that a transcriptional function of YAP was repressing Wnt, we generated $Yap^{fl/S79A}$ *Villin-Cre* mice. The YAP protein in these mutants cannot bind to TEAD transcription factors, the main transcriptional effectors of YAP[18,19]. After Rspo1 injection, we observed no enhanced Wnt response in YAP(S79A) mutant mice (Supplementary Fig. 7a). Supporting a role for cytoplasmic YAP in restricting Wnt signalling, expression of YAP wild-type and a YAP(S127D) phospho-mimic limited Wnt reporter responsiveness in 293T cells (Supplementary Fig. 7b, c).

The phenotype observed in cKO mice treated with Rspo1 histologically resembled acute *Apc* deletion (Supplementary Fig. 8)[20]. Surprisingly, we observed no obvious changes in β-catenin protein levels in hyperplastic cKO crypts (Supplementary Fig. 8a–c). As increases in β-catenin protein are the direct consequence of disrupting the AXIN–APC–GSK-3β complex, these data suggest that YAP restriction of Wnt signalling is probably not mediated by modulating the activity of this complex. It has been recently suggested that phosphorylated cytoplasmic YAP sequesters β-catenin in the cytoplasm in cell lines[21]. Although we observed a subtle expansion in the number of nuclear β-catenin-positive cells in cKO mucosa (Supplementary Fig. 8b), these probably represent the expansion of Paneth cells and we infer that this does not represent the major mechanism of YAP-mediated Wnt repression. *In vitro*, we observed increased Wnt activity after combined YAP and APC depletion versus knockdown of APC only, and a synergistic effect of YAP depletion together with GSK-3β small-molecule inhibition (Supplementary Fig. 8d, e). In mice, combined deletion of APC and YAP (double knockout) versus APC only in an acute or a long-term adenoma model[22] resulted in a robust upregulation of CD44 in addition to significant increases in the number of Paneth cells and proliferative cells (Supplementary Fig. 9a–g). Thus, our data suggest that YAP acts to restrict Wnt signalling independently to, and in some cases synergistically with, the activity of the destruction complex and control of subcellular localization of β-catenin.

Recently, it was reported that TAZ, a protein that bears a 57% homology to YAP, interacts with DVL[23], a positive Wnt signalling regulator. Work primarily done in *Drosophila* has shown that DVL acts in the cytoplasm upstream of the β-catenin destruction complex. However, emerging work in mammals suggests that DVL also exists in the nucleus, mediating the Wnt transcriptional response partly in conjunction with c-Jun and TCF4 (refs 24, 25). We determined that YAP also interacts with DVL2 (Supplementary Fig. 10a). The function of DVLs in intestinal regeneration has been poorly characterized, although the observation that loss of DVL2 causes a decrease in intestine length and crypt diameter suggests an important role in

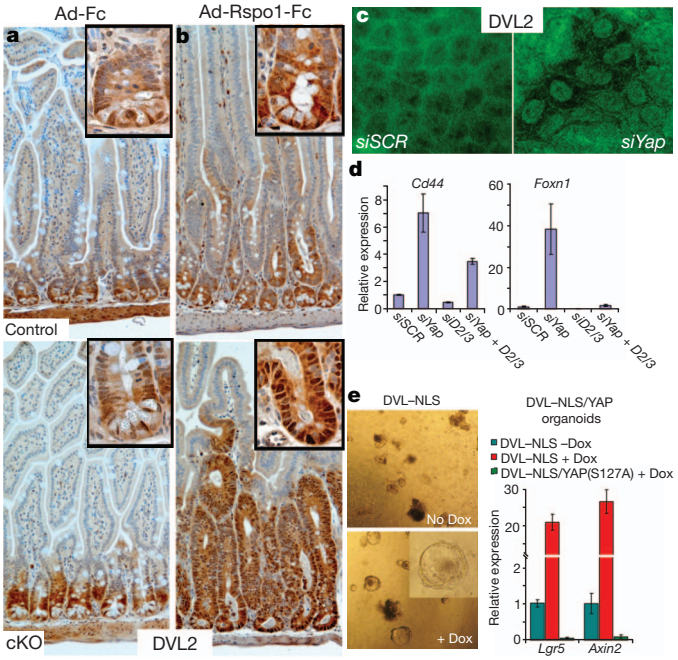


Figure 3 | YAP restricts Wnt signalling by blocking DVL nuclear translocation. **a**, **b**, DVL2 immunohistochemistry in control and cKO small intestine treated with Ad-Fc (**a**) and Ad-Rspo1-Fc (**b**). Insets are higher magnifications showing subcellular localization of DVL2. **c**, Immunofluorescent staining for DVL2 in confluent DLD1 cells after transfection with indicated short interfering RNA (siRNA). siSCR, scrambled siRNA. **d**, Expression analysis of Wnt target genes in DLD1 cells after transfection with siRNAs against YAP, DVL2 plus DVL3 (D2/3) or YAP plus DVL2/3. **e**, Expression analysis of organoids infected with lentivirus encoding Dox-inducible DVL–NLS alone, or on a Villin-rtTA TetO-YAP(S127A) background. Dox was given in culture medium for 4 days. All graphed data represent the mean and standard deviation of triplicate cultures. Original magnifications in each panel are ×10 (**a**, **b**), ×20 (**c**) and ×4 (**e**).

intestinal biology[26]. We find that DVL2 expression is restricted to the crypt compartment, where it is localized in the nucleus of CBCs and assumes a more diffuse pattern in the TA compartment (Fig. 3a). Furthermore, DVL2 nuclear localization is enhanced in crypts after Rspo1 administration and during regeneration following irradiation (Fig. 3b and Supplementary Fig. 10b). To evaluate the role of nuclear DVL in crypts, we infected intestinal organoids with lentiviruses expressing a Dox-inducible DVL carrying a nuclear localization signal (DVL–NLS)[25]. DVL–NLS increases Wnt target gene expression and leads to the transient formation of spheroid organoids resembling those with constitutive Wnt signalling (Fig. 3e)[9]. In addition, expression of DVL–NLS after APC knockdown results in a synergistic activation of the TOPflash Wnt reporter (Supplementary Fig. 10c). Thus, our findings are consistent with an important nuclear function of DVL in intestinal progenitors and Wnt signalling in parallel to or downstream of the destruction complex.

Given the observed physical interaction between YAP and DVL2 and their overlapping patterns of subcellular localization in crypts, we proposed that YAP might control DVL2 subcellular localization and/or activity. We observed a massive expansion of cells positive for nuclear DVL2 in cKO mice treated with Rspo1 and during radiation-induced regeneration (Fig. 3b and Supplementary Fig. 10b). Additionally, YAP knockdown led to DVL nuclear accumulation in DLD1 cells (Fig. 3c and Supplementary Fig. 10d, e). We found no increase in levels of nuclear β-catenin in this context (Supplementary Fig. 10e). Loss of DVL2 and DVL3 partially or completely rescued increased Wnt target gene expression in the absence of YAP (Fig. 3d). To determine if YAP could block DVL-induced Wnt transcriptional responses, we generated organoids expressing Dox-inducible DVL–NLS (Supplementary Fig. 10f) and YAP. Expression of YAP completely abrogated the DVL-mediated upregulation of the Wnt target genes *Lgr5* and *Axin2* (Fig. 3e). Therefore, we conclude that YAP is critical for the proper subcellular localization of DVLs and blocks their ability to induce Wnt signalling.

Finally, we investigated the role of YAP in human colorectal carcinoma. Using DLD1 cells infected with Dox-inducible YAP wild type or YAP(S127D), we performed xenograft assays to assess the effect of YAP on tumour growth. We found a marked decrease in tumour growth with the addition of Dox, particularly with YAP(S127D) (Fig. 4a and Supplementary Fig. 11a, b). This decrease in tumour size coincided with global suppression of the colorectal carcinoma TCF4/β-catenin[27] and ISC[11] gene signatures (Fig. 4b, c, Supplementary Fig. 11c, d and Supplementary Table 1). We next examined if YAP loss was associated with human colorectal carcinoma by evaluating YAP expression in a cohort of 672 colorectal carcinoma samples using immunohistochemistry[28]. Complete loss of YAP staining occurred within a minor subset of human tumours (10.5%), but predicted worse patient survival and was associated with high-grade, stage IV disease, compared with YAP-positive groups (Supplementary Fig. 11e–h and Fig. 4d–f). These results are consistent with our animal data and demonstrate that YAP may act as a tumour suppressor in human colorectal carcinoma.

Precise mechanisms must exist to dampen proliferative potential to maintain organ size. We have identified YAP and DVLs as key players in this process by restricting the expansion of ISCs and their niche during Wnt-driven regeneration. Although we have not ruled out a pro-proliferative role for nuclear YAP, our results suggest that the predominant function of YAP in the intestine is growth restrictive. During normal development and homeostasis, TAZ (whose expression is highly enriched in crypts; Supplementary Fig. 6d) might be compensating for the loss of YAP. Alternatively, YAP together with DVL might be part of a regeneration-exclusive molecular machinery that helps counteract hyperactive Wnt signalling, which occurs during emergency growth. Interestingly, *YAP* is a WNT target gene[29], suggesting that it might participate in a negative feedback loop to limit WNT-initiated signals. At present, therapeutic efforts are aimed at diminishing YAP protein levels in many malignancies, including colorectal carcinoma[30]. Our results suggest that such manipulations could potentially result in increased tumour growth. YAP-targeting therapies should aim to disrupt the YAP–TEAD interaction or to induce the cytoplasmic translocation of YAP. Further study of the role of YAP as a growth suppressor is likely to provide important insights into the biology of regeneration, size control and tumour progression.

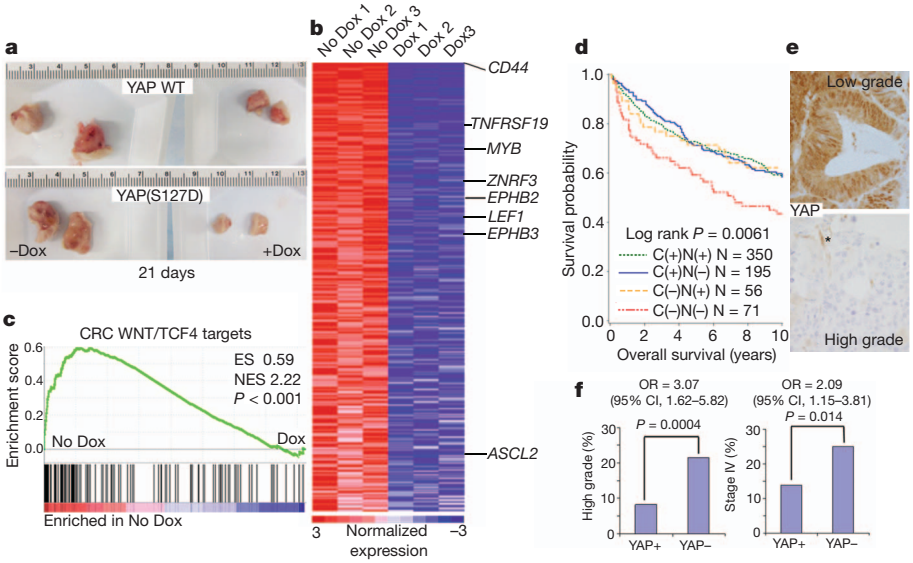


Figure 4 | YAP function in human colorectal cancer. **a**, Xenograft tumour formation of DLD1 cells infected with lentivirus encoding Dox-inducible YAP wild type (WT) or YAP(S127D) mutant. Tumour photographs are representative images. Ruler is in centimetres. **b**, Heatmap representing the top 328 genes downregulated by YAP(S127D) induction in xenograft assays ($n = 3$ tumours per treatment group). Labelled genes are known β-catenin targets in human colorectal carcinoma. **c**, GSEA of TCF4-dependent colorectal carcinoma (CRC) target genes. **d**, Overall survival analysis of patients with specific YAP staining patterns. YAP immunohistochemistry was placed into four groups of staining patterns correlating with subcellular localization from 672 colorectal cancer patients: cytoplasmic and nuclear (C (+) N (+)); cytoplasmic (C (+) N (−)); nuclear positive (C (−) N (+)); and complete loss of staining (C (−) N (−)). **e**, YAP staining in low- and high-grade tumours. Original magnification, ×20. **f**, YAP loss is significantly associated with high-grade tumours and stage IV disease (CI, confidence interval; OR, odds ratio).

METHODS SUMMARY

Immunohistochemistry. Staining was performed with the following antibodies using the Vectastain ABC-Elite kit according to the manufacturer's instructions, except for Cryptdin1, which was detected using an anti-goat Alexa-Fluor 488 (Invitrogen). Rabbit anti-YAP (Cell Signaling 1:40, Avruch Lab 1:400), mouse anti-PCNA (Santa Cruz Biotechnology), rat anti-CD44 (BD Biosciences 1:100), rat anti-Ki67 (Dako 1:40), rabbit anti-phospho-histone H3 (Millipore 1:300), rabbit anti-SOX9 (Millipore 1:300), goat anti-EPHB3 (R&D 1:300), rabbit anti-GFP (Abcam 1:3,000), goat anti-Cryptdin1 (A. Ouellette), rabbit anti-DVL2 (Cell Signaling 1:100).

Full Methods and any associated references are available in the online version of the paper.

Received 15 May; accepted 19 October 2012.
Published online 25 November 2012.

1. Pan, D. The Hippo signaling pathway in development and cancer. *Dev. Cell* **19,** 491–505 (2010).
2. Ramos, A. & Camargo, F. D. The Hippo signaling pathway and stem cell biology. *Trends Cell Biol.* **22,** 339–346 (2012).
3. Camargo, F. D. *et al.* YAP1 increases organ size and expands undifferentiated progenitor cells. *Curr. Biol.* **17,** 2054–2060 (2007).
4. Dong, J. *et al.* Elucidation of a universal size-control mechanism in *Drosophila* and mammals. *Cell* **130,** 1120–1133 (2007).
5. Zhang, J. *et al.* YAP-dependent induction of amphiregulin identifies a non-cell-autonomous component of the Hippo pathway. *Nature Cell Biol.* **11,** 1444–1450 (2009).
6. Roth, S. *et al.* Generation of a tightly regulated doxycycline-inducible model for studying mouse intestinal biology. *Genesis* **47,** 7–13 (2009).
7. Pinto, D., Gregorieff, A., Begthel, H. & Clevers, H. Canonical Wnt signals are essential for homeostasis of the intestinal epithelium. *Genes Dev.* **17,** 1709–1713 (2003).
8. Andreu, P. *et al.* A genetic study of the role of the Wnt/β-catenin signalling in Paneth cell differentiation. *Dev. Biol.* **324,** 288–296 (2008).
9. Sato, T. *et al.* Paneth cells constitute the niche for Lgr5 stem cells in intestinal crypts. *Nature* **469,** 415–418 (2011).
10. van der Flier, L. G. *et al.* Transcription factor achaete scute-like 2 controls intestinal stem cell fate. *Cell* **136,** 903–912 (2009).
11. Muñoz, J. *et al.* The Lgr5 intestinal stem cell signature: robust expression of proposed quiescent '+4' cell markers. *EMBO J.* **31,** 3079–3091 (2012).
12. Fevr, T., Robine, S., Louvard, D. & Huelsken, J. Wnt/β-catenin is essential for intestinal homeostasis and maintenance of intestinal stem cells. *Mol. Cell. Biol.* **27,** 7551–7559 (2007).
13. Cai, J. *et al.* The Hippo signaling pathway restricts the oncogenic potential of an intestinal regeneration program. *Genes Dev.* **24,** 2383–2388 (2010).
14. Ashton, G. H. *et al.* Focal adhesion kinase is required for intestinal regeneration and tumorigenesis downstream of Wnt/c-Myc signaling. *Dev. Cell* **19,** 259–269 (2010).
15. Davies, P. S., Dismuke, A. D., Powell, A. E., Carroll, K. H. & Wong, M. H. Wnt-reporter expression pattern in the mouse intestine during homeostasis. *BMC Gastroenterol.* **8,** 57 (2008).
16. Ootani, A. *et al.* Sustained *in vitro* intestinal epithelial culture within a Wnt-dependent stem cell niche. *Nature Med.* **15,** 701–706 (2009).
17. Kim, K. A. *et al.* Mitogenic influence of human R-spondin1 on the intestinal epithelium. *Science* **309,** 1256–1259 (2005).
18. Zhao, B. *et al.* TEAD mediates YAP-dependent gene induction and growth control. *Genes Dev.* **22,** 1962–1971 (2008).
19. Schlegelmilch, K. *et al.* Yap1 acts downstream of α-catenin to control epidermal proliferation. *Cell* **144,** 782–795 (2011).
20. Sansom, O. J. *et al.* Loss of Apc *in vivo* immediately perturbs Wnt signaling, differentiation, and migration. *Genes Dev.* **18,** 1385–1390 (2004).
21. Imajo, M., Miyatake, K., Iimura, A., Miyamoto, A. & Nishida, E. A molecular mechanism that links Hippo signalling to the inhibition of Wnt/β-catenin signalling. *EMBO J.* **31,** 1109–1122 (2012).
22. Cheung, A. F. *et al.* Complete deletion of Apc results in severe polyposis in mice. *Oncogene* **29,** 1857–1864 (2010).
23. Varelas, X. *et al.* The Hippo pathway regulates Wnt/β-catenin signaling. *Dev. Cell* **18,** 579–591 (2010).
24. Itoh, K., Brott, B. K., Bae, G. U., Ratcliffe, M. J. & Sokol, S. Y. Nuclear localization is required for Dishevelled function in Wnt/β-catenin signaling. *J. Biol.* **4,** 3 (2005).
25. Gan, X. Q. *et al.* Nuclear Dvl, c-Jun, β-catenin, and TCF form a complex leading to stabilization of β-catenin–TCF interaction. *J. Cell Biol.* **180,** 1087–1100 (2008).
26. Metcalfe, C. *et al.* Dvl2 promotes intestinal length and neoplasia in the Apc^{Min} mouse model for colorectal cancer. *Cancer Res.* **70,** 6629–6638 (2010).
27. Van der Flier, L. G. *et al.* The intestinal Wnt/TCF signature. *Gastroenterology* **132,** 628–632 (2007).
28. Chan, A. T., Ogino, S. & Fuchs, C. S. Aspirin and the risk of colorectal cancer in relation to the expression of COX-2. *N. Engl. J. Med.* **356,** 2131–2142 (2007).
29. Konsavage, W. J., Kyler, S. L., Rennoll, S. A., Jin, G. & Yochum, G. S. Wnt/β-catenin signaling regulates Yes-associated protein (YAP) gene expression in colorectal carcinoma cells. *J. Biol. Chem.* **287,** 11730–11739 (2012).
30. Avruch, J., Zhou, D. & Bardeesy, N. YAP oncogene overexpression supercharges colon cancer proliferation. *Cell Cycle* **11,** 1090–1096 (2012).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank L. Zon and Camargo laboratory members for critical review of this manuscript and X. He for Myc–DVL2 constructs. This work was supported by grants from the Stand Up to Cancer-AACR initiative (F.D.C.), National Institutes of Health R01 CA131426 and AR064036 (F.D.C.) and the Harvard Stem Cell Institute (F.D.C.). F.D.C. is a Pew Scholar in the Biomedical Sciences. E.R.B. is supported by a postdoctoral fellowship from the American Cancer Society Illinois Division (PF-12-245-01-CCG). This work has also benefited from support to K.S.Y. (CIRM, and 1K08DK096048), C.J.K. (1U01DK085527), S.O. (R01 CA151993 and R01 DK091427), S.T.M. (P01CA87969) and C.S.F. (P50CA127003).

Author Contributions E.R.B. and F.D.C. designed the study. E.R.B., B.L.B., K.S., R.d.I.R. and S.T.M. performed experiments. T.M. performed immunohistochemistry and analysis of human tumour samples. K.S.Y., C.S.F., R.S., S.O. and C.J.K. provided reagents for the study.

Author Information Microarray data have been deposited in the Gene Expression Omnibus database under accession number GSE41509. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to F.D.C. (Fernando.camargo@childrens.harvard.edu).

METHODS

Mice. *Yap* conditional knockout and S79A point mutant mice were previously described[19]. *Yap* cKO mice were intercrossed with *Villin-Cre*[31], *Villin-CreER*[32], *Lgr5-EGFP-IRES-CreER*[33] and *Apc*$^{fl/fl}$ mice[34]. Mice were given a single dose of tamoxifen (1 mg in corn oil injected intraperitoneally) in *Yap/Apc* conditional knockout experiments with *Villin-CreER*. *Villin-rtTA* mice were described previously[6].

For *in vivo* administration of R-Spondin1, adenovirus expressing Fc-tagged R-Spondin1 of mouse origin was retro-orbitally injected at 3×10^8 plaque-forming units (p.f.u.). Virus was produced as previously described[16]. Tissue was collected at 3.5 or 7 days. For irradiation experiments, mice were exposed to 9 or 11 Gy at a dose of 0.664 Gy min^{-1} from a caesium 137 irradiator. Tissue was collected 5 days after irradiation.

Immunohistochemistry of mouse tissue. Staining was performed with the following antibodies using the Vectastain ABC-Elite kit according to the manufacturer's instructions, with the exception of Cryptdin1, which was detected using an anti-goat Alexa-Fluor 488 (Invitrogen). Rabbit anti-YAP (Cell Signaling 1:40, Avruch Lab 1:400), mouse anti-PCNA (Santa Cruz Biotechnology), rat anti-CD44 (BD Biosciences 1:100), rat anti-Ki67 (Dako 1:40), rabbit anti-phospho-histone H3 (Millipore 1:300), rabbit anti-SOX9 (Millipore 1:300), goat anti-EPHB3 (R&D 1:300), rabbit anti-GFP (Abcam 1:3,000), goat anti-Cryptdin1 (Andre Ouellette), rabbit anti-DVL2 (Cell Signaling 1:100). Antigen retrieval was performed using low pH citrate buffer. Staining for β-catenin was performed as previously described[35]. Lysozyme (Dako rabbit anti-human lysozyme) staining was performed as follows. Antigen retrieval was performed by proteinase-K digestion at 37 °C for exactly 12 min in tris-EDTA pH 9 followed by PBS wash for 5 min and quenching of endogenous peroxidases and Vectastain ABC-Elite kit processing. Next, tissue was incubated in primary antibody at 1:1,000 in PBS 0.1% Triton at 4 °C overnight.

ISH. ISH for *Olfm4* transcripts was performed on formalin-fixed intestine at the specialized histopathology service core at the Dana Farber Cancer Institute using RNAscope, according to the manufacturer's instructions.

Quantification of phosphorylated H3S10 and lysozyme foci. For Villin-CreER experiments, lysozyme quantification was performed on a cohort of five control (*Apc*$^{fl/fl}$) and five double-cKO (*Yap*$^{fl/fl}$ *Apc*$^{fl/fl}$) mice. The quantification was performed blind to specimen identity. Paneth cell numbers were quantified by the number of crypt cells staining positive for lysozyme on a per crypt basis. For each sample, per crypt numbers of lysozyme-positive cells were averaged over at least 15 ×10 fields of view per mouse. In Villin-CreSt experiments, phosphorylated H3S10 and lysozyme positivity was counted in a similar manner. Student's *t*-test was used to determine statistical significance with a cut-off of $P < 0.05$.

DSS treatment. Mice were given 2.5% DSS in drinking water for 7 days, followed by normal drinking water for 3 days. Colon tissue was snap frozen and cryosections were stained with haematoxylin to visualize tissue architecture.

siRNA-mediated knockdown. Cells were reverse transfected with a 7 nM final concentration of the indicated siRNA (Ambion Silencer Select) using Lipofectamine RNAiMax (Invitrogen), according to the manufacturer's instructions. For expression analysis in DLD1 cells, RNA was extracted 4 days after transfection. Oligonucleotides against human transcripts used were (5′-3′): *siAPC* sense, GG AUCUGUAUCAAGCCGUUTT; antisense, AACGGCUUGAUACAGAUCCTT (oligonucleotide no. s1433); *siYAP* sense, GGUGAUACUAUCAACCAAATT; antisense, UUUGGUUGAUGUAUCACCTG (oligonucleotide no. s20366); *siDVL3* sense, GAUAUGUUGUUACAGGUAATT; antisense, UUACCUGUAACAAC AUAUCTC (oligonucleotide no. s675); *siDVL2* (pooled two siRNAs) sense, CA CCAUCCCUAAUGCCUUUTT; antisense, AAAGGCAUUAGGGAUGGUGAT (oligonucleotide no s4396); and sense, CAGUCACGCUAAACAUGGATT; antisense, UCCAUGUUUAGCGUGACUGTG (oligonucleotide no. s4398).

RNA extraction, cDNA synthesis and qPCR. RNA was extracted from cells or tissue using an RNeasy mini kit according to the manufacturer's instructions (Qiagen). cDNA synthesis was performed on 1 μg of total RNA using iScript cDNA synthesis kit (Biorad). qPCR was performed using Taqman probes available from Applied Biosystems and Taqman Fast-Advanced master mix. Probed used were as follows: mouse *Cd44* Mm01277163_m1; *Lgr5* Mm00438890_m1; *Myc* Mm004877804; *Wwtr1* (*Taz*) Mm00513560_m1; *Nkd1* Mm00471902_m1; *Yap* Mm00494240_m1; *Sox9* Mm00448840_m1; *Cd133* (*Prom1*) Mm00477115_m1; *Axin2* Mm00443610_m1; *Apc* Mm00545877_m1; and *Fabp2* Mm00433188_m1. Human probes were: *NKD1* Hs00263894_m1 and *FOXN1* Hs00263894_m1. All reactions were run on Applied Biosystems Step-One-Plus real-time PCR instrument and data were calculated using the $\Delta\Delta C_t$ method. Significance values were calculated using Student's *t*-test with a *P*-value cut-off of less than 0.05.

Immunofluorescence. For immunocytochemistry of adherent DLD1 cells, cultures were grown on sterile coverslips and washed once with PBS followed by fixation in 4% paraformaldehyde/PBS for 15 min at room temperature (~20 °C). Cells were then washed twice with cold PBS and permeabilized with PBS/0.25% Triton X-100 for 10 min at room temperature. Cells were then washed three times in PBS followed by blocking in 1% BSA/PBS for 30 min at room temperature. Next, cells were incubated in primary antibodies in 1% BSA/PBS overnight at 4 °C followed by 3 washes in PBS. Cells were then incubated for 1 h in a dark humidified chamber with Alexa-Fluor-conjugated secondary antibodies in 1% BSA/PBS followed by 3 washes in PBS. Coverslips were then stained with 4′,6-diamidino-2-phenylindole (DAPI) and mounted on glass slides. Staining was performed with the following antibodies: mouse anti-YAP (Santa Cruz 1:500) and rabbit anti-DVL2 (Millipore 1:200).

Co-immunoprecipitation. 293T cells were transiently transfected with GFP (negative control) or indicated Flag-tagged versions of YAP2 (Plasmids purchased from Addgene, originally constructed in the laboratory of M. Sudol) or Myc-tagged DVL2 (from the X. He lab) using Lipofectamine 2000. After 48 h, cells were lysed in buffer containing 1% Triton X-100, 50 mM Tris HCl pH 7.4, 150 mM NaCl and 1 mM EDTA followed by passage through a 26-gauge needle ten times. Lysates were pre-cleared with protein-A agarose and then incubated in anti-Flag resin (Sigma Aldrich) for 1 h 30 min while shaking. Protein complexes were then washed four times in lysis buffer and boiled for western detection. Mouse anti-Flag M2 antibody (Sigma Aldrich) was used at 1:1,000 to detect Flag-tagged YAP and rabbit anti-Myc antibody (Sigma Aldrich 1:1,000) was used to detect Myc-tagged DVL2.

Subcellular fractionation. DLD1 cells were reverse transfected with the indicated siRNA. After 4 days, cells were washed with PBS, scraped and centrifuged. Cells were then resuspended in 3–4 pellet volumes of fractionation buffer containing sucrose, HEPES, KCl, MgCl, EDTA, NP-40, protease and phosphatase inhibitors. Cells were then passed through a 25.5-gauge needle ten times and left on ice for 20 min and then centrifuged to pellet nuclei away from cytoplasm. Nuclei were then washed by passage through a 25.5-gauge needle ten times and centrifuged. Nuclear and cytoplasmic fractions were then resuspended in equal volumes of lysis buffer and analysed by western blots. The following antibodies were used: mouse anti-YAP (Santa Cruz 1:1,000), rabbit anti-DVL2 (Cell Signaling 1:1,000), anti-fibrillarin (Abcam 1:1,000), anti-β-tubulin (Cell Signaling 1:1,000).

Wnt reporter assays. 293T cells were first transfected with the indicated siRNA. Two days later, TOPflash and renilla plasmids were cotransfected. Wnt activity was assayed 48 h after reporter transfection. All values are represented as the ratio of firefly to renilla. Transfections were performed in triplicate.

A stable 293T cell line was used in BIO and Wnt3A plus Rspo1 induction experiments. This line was created by co-transfecting the TOPflash plasmid with a renilla-expressing plasmid (10:1 ratio of firefly to renilla) carrying a blasticidin resistance selection marker. Cells were selected under 6 μg ml^{-1}. Clones were picked, expanded and screened for TOPflash activity after treatment with recombinant Wnt3A. YAP and scrambled siRNAs were transfected at 7 nM, and induced with Wnt3A and Wnt3A (500 ng ml^{-1}) plus RSpondin1 (500 ng ml^{-1}) 3 days later when cells were approximately 50% confluent, for 16 h. Induction with bromoindirubin (5 μM) (BIO Sigma Aldrich) was performed in the same manner as with Wnt3A and Rspo1. Controls were treated with DMSO. To test the effect of nuclear localized DVL in 293T WNT reporter assays, a DVL–NLS-expressing construct[25] was transiently transfected with renilla and TOPflash plasmids 3 days after siRNA treatment. Luciferase levels were measured 48 h later.

Microarray analysis. In transgenic experiments, whole RNA was extracted from mouse crypts. RNA was extracted from whole epithelium in Rspo1 experiments. Microarray experiments were performed by the Microarray Core of the Molecular Genetics Core Facility at the Children's Hospital Boston. Material was processed for Affymetrix Mouse GeneChip 1.0ST arrays. For each experimental condition, three mice were used. CEL files were converted to GCT files using ExpressionFileCreator in Gene Pattern, publicly available from the Broad Institute (http://www.genepattern.broadinstitute.org). GCT files were preprocessed and used in the Comparative Marker Selection module to obtain lists of differentially regulated genes. In transgenic experiments, a threshold of $P \leq 0.05$ and fold change of 1.4 or greater was used. In Rspo1 experiments, a fold change of 2 and *q* value (FDR-corrected *P* value) of 0.05 were used as cut-offs. Expression analysis of human xenograft tumours was performed using Affymetrix Human GeneChip 1.0ST chips and processed as above. GCT files were used in Comparative Marker Selection. RNA was used from three control (no Dox) and three YAP(S127D) Dox-induced tumours. Threshold cut-offs were $P \leq 0.05$ and a fold change of 1.4-fold or greater. GSEA was performed using GCT files from Yap experiments compared to gene sets cited in the main text.

Organoid culture and lentiviral infection. Organoids were derived, propagated and infected according to previously published work[36,37]. pINDUCER20[38] was used for inducible lentiviral expression studies in DLD1 xenografts and organoids.

Human study population. The databases of two nationwide prospective cohort studies were used: the Nurses' Health Study ($N = 121,701$ women followed since 1976); and the Health Professionals Follow-up Study ($N = 51,529$ men followed since 1986)[28]. We collected paraffin-embedded tissue blocks from hospitals

throughout the United States where patients underwent colorectal cancer resections. Hematoxylin and eosin stained tissue sections from all colorectal cancer cases were reviewed by a pathologist (S.O.) unaware of other data. Tumour grade was categorized as low versus high (>50% versus ≤ 50% gland formation). We excluded cases that were preoperatively treated. Patients were observed until death or until January 1, 2011, whichever came first. Death of a participant was confirmed by the National Death Index. Returning a questionnaire indicated informed consent from all study participants. This study was approved by the Human Subjects Committees at Harvard School of Public Health and Brigham and Women's Hospital.

Immunohistochemistry on human samples. Tissue microarrays were constructed as previously described[39]. For YAP immunohistochemistry, deparaffinized tissue sections were treated with Antigen Retrieval Citra Solution (Biogenex Laboratories) in a microwave for 15 min. Tissue sections were then incubated with Peroxidase Blocking Reagent (15 min; DAKO) and primary antibody against YAP (rabbit polyclonal, 1:400 dilution; Cell Signaling) was applied, and slides were incubated for 16 h at 4 °C. Next, we applied SignalStain Boost IHC Detection Reagent (Cell Signaling) for 30 min followed by visualizing signal with diaminobenzidine (5 min; DAKO) and haematoxylin counterstain. Each immunohistochemical maker was evaluated by a pathologist (T.M.) unaware of other data.

Statistical analysis. All statistical analyses were performed with the SAS program (version 9.1, SAS Institute). All *P* values were two-sided and statistical significance was set at $P = 0.05$. For categorical data, the chi-square test was performed. Kaplan–Meier method and log-rank test were used for survival analyses.

31. Madison, B. B. *et al. cis* elements of the villin gene control expression in restricted domains of the vertical (crypt) and horizontal (duodenum, cecum) axes of the intestine. *J. Biol. Chem.* **277,** 33275–33283 (2002).
32. El Marjou, F. *et al.* Tissue-specific and inducible Cre-mediated recombination in the gut epithelium. *Genesis* **39,** 186–193 (2004).
33. Barker, N. *et al.* Identification of stem cells in small intestine and colon by marker gene *Lgr5*. *Nature* **449,** 1003–1007 (2007).
34. Colnot, S. *et al.* Colorectal cancers in a new mouse model of familial adenomatous polyposis: influence of genetic and environmental modifiers. *Lab. Invest.* **84,** 1619–1630 (2004).
35. Barker, N. *et al.* Crypt stem cells as the cells-of-origin of intestinal cancer. *Nature* **457,** 608–611 (2009).
36. Koo, B. K. *et al.* Controlled gene expression in primary *Lgr5* organoid cultures. *Nature Methods* **9,** 81–83 (2012).
37. Sato, T. *et al.* Single Lgr5 stem cells build crypt–villus structures *in vitro* without a mesenchymal niche. *Nature* **459,** 262–265 (2009).
38. Meerbrey, K. L. *et al.* The pINDUCER lentiviral toolkit for inducible RNA interference *in vitro* and *in vivo*. *Proc. Natl Acad. Sci. USA* **108,** 3665–3670 (2011).
39. Ogino, S. *et al.* Combined analysis of COX-2 and p53 expressions reveals synergistic inverse correlations with microsatellite instability and CpG island methylator phenotype in colorectal cancer. *Neoplasia* **8,** 458–464 (2006).

LETTER

doi:10.1038/nature11699

Ca^{2+} regulates T-cell receptor activation by modulating the charge property of lipids

Xiaoshan Shi[1,2]*, Yunchen Bi[3]*, Wei Yang[1,2]*, Xingdong Guo[1,2], Yan Jiang[1,2], Chanjuan Wan[3], Lunyi Li[1,2], Yibing Bai[1,2], Jun Guo[1,2], Yujuan Wang[3], Xiangjun Chen[4], Bo Wu[3], Hongbin Sun[3], Wanli Liu[4], Junfeng Wang[3] & Chenqi Xu[1,2]

Ionic protein–lipid interactions are critical for the structure and function of membrane receptors, ion channels, integrins and many other proteins[1–7]. However, the regulatory mechanism of these interactions is largely unknown. Here we show that Ca^{2+} can bind directly to anionic phospholipids and thus modulate membrane protein function. The activation of T-cell antigen receptor–CD3 complex (TCR), a key membrane receptor for adaptive immunity, is regulated by ionic interactions between positively charged CD3ε/ζ cytoplasmic domains ($CD3_{CD}$) and negatively charged phospholipids in the plasma membrane[1,8–10]. Crucial tyrosines are buried in the membrane and are largely protected from phosphorylation in resting T cells. It is not clear how $CD3_{CD}$ dissociates from the membrane in antigen-stimulated T cells. The antigen engagement of even a single TCR triggers a Ca^{2+} influx[11] and TCR-proximal Ca^{2+} concentration is higher than the average cytosolic Ca^{2+} concentration[12]. Our biochemical, live-cell fluorescence resonance energy transfer and NMR experiments showed that an increase in Ca^{2+} concentration induced the dissociation of $CD3_{CD}$ from the membrane and the solvent exposure of tyrosine residues. As a consequence, CD3 tyrosine phosphorylation was significantly enhanced by Ca^{2+} influx. Moreover, when compared with wild-type cells, Ca^{2+} channel-deficient T cells had substantially lower levels of CD3 phosphorylation after stimulation. The effect of Ca^{2+} on facilitating CD3 phosphorylation is primarily due to the charge of this ion, as demonstrated by the fact that replacing Ca^{2+} with the non-physiological ion Sr^{2+} resulted in the same feedback effect. Finally, ^{31}P NMR spectroscopy showed that Ca^{2+} bound to the phosphate group in anionic phospholipids at physiological concentrations, thus neutralizing the negative charge of phospholipids. Rather than initiating CD3 phosphorylation, this regulatory pathway of Ca^{2+} has a positive feedback effect on amplifying and sustaining CD3 phosphorylation and should enhance T-cell sensitivity to foreign antigens. Our study thus provides a new regulatory mechanism of Ca^{2+} to T-cell activation involving direct lipid manipulation.

TCR is one of the most complicated cell surface receptors and is composed of a ligand-sensing TCRαβ subunit and the three signalling subunits CD3εδ, εγ and ζζ. TCR recognition of the peptide–major histocompatibility complex (pMHC) is an essential step in initiating the host adaptive immune response against invading pathogens[13]. Antigen binding first triggers the tyrosine phosphorylation of the immunoreceptor tyrosine-based activation motifs (ITAMs) in CD3 chains, which is primarily mediated by Lck[14], and consequently elicits a comprehensive signalling network hallmarked by Ca^{2+} influx through the Ca^{2+}-release-activated Ca^{2+} channel (CRAC)[15,16]. The regulatory mechanism of ITAM phosphorylation is a longstanding puzzle[17]. Another central question that remains is how T cells can be activated by minute quantities of foreign antigen[11,18]. Recent work shows that ITAM phosphorylation is regulated by the ionic interaction between the positively charged ITAM and the negatively charged phospholipids in the plasma membrane[1,8–10,19,20]. In resting T cells, the CD3ε and ζ cytoplasmic domains bind to the plasma membrane, and their ITAM tyrosines are inserted into the hydrophobic core of the plasma membrane[1,8–10]. The insertion prevents spontaneous ITAM phosphorylation

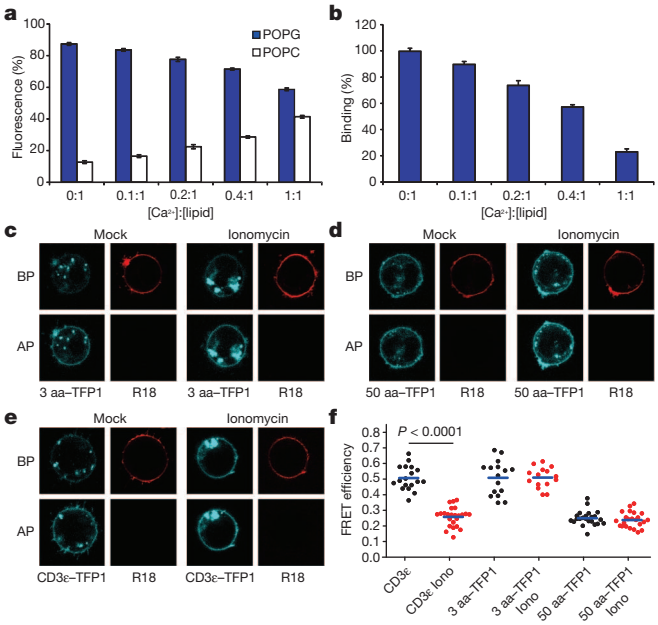


Figure 1 | Ca^{2+} induced the dissociation of $CD3\varepsilon_{CD}$ from the membrane bilayer. **a**–**f**, The effect of Ca^{2+} on the $CD3\varepsilon_{CD}$–membrane interaction was measured by an equilibrium-based microdialysis assay (**a**, **b**) and a live-cell FRET imaging experiment (**c**–**f**). **a**, In the absence of Ca^{2+}, $CD3\varepsilon_{CD}$ bound strongly to POPG large unilamellar vesicles, not to POPC large unilamellar vesicles. The increase of Ca^{2+} concentration significantly impaired the $CD3\varepsilon_{CD}$–POPG interaction. Triplicate measurements were made for each condition and the results (mean ± s.d.) were plotted as the percentage of the fluorescence intensity in the experimental (POPG) and control (POPC) chambers. **b**, The results in **a** were converted to the binding efficiency of $CD3\varepsilon_{CD}$ to POPG and presented as mean ± s.d. **c**–**f**, The FRET efficiency was measured between TFP1 and the membrane dye R18 by the dequenching approach in live Jurkat T cells. Ionomycin (5 μM) was used to induce Ca^{2+} influx. The FRET efficiencies of the 3 aa–TFP1 and 50 aa–TFP1 control constructs were not sensitive to Ca^{2+} influx, whereas the FRET efficiency of $CD3\varepsilon_{CD}$–TFP1 was reduced significantly in response to Ca^{2+} influx. AP, after photobleaching; BP, before photobleaching. The FRET efficiencies were measured for more than 15 cells per condition. Each dot represents the FRET value from one individual cell. Unpaired two tailed Student's *t*-test was used for the statistical analysis.

[1]State Key Laboratory of Molecular Biology, Institute of Biochemistry and Cell Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, 320 Yueyang Road, Shanghai 200031, China. [2]National Center for Protein Science Shanghai, Institute of Biochemistry and Cell Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, 333 Haike Road, Shanghai 201203, China. [3]High Magnetic Field Laboratory, Hefei Institutes of Physical Science, Chinese Academy of Sciences, 350 Shushanhu Road, Hefei 230031, Anhui Province, China. [4]School of Life Sciences, Tsinghua University, Beijing 100084, China.
*These authors contributed equally to this work.

in unstimulated T cells even when active Lck proteins are constitutively available[21]. In antigen-stimulated T cells, ITAMs need to be dislodged from the plasma membrane and become accessible for Lck. Immediately after the initial TCR triggering, Ca^{2+} locally influxes into T cells[22]. Ca^{2+} ions are not evenly distributed in T cells and we found that Ca^{2+} microdomains co-localized with TCR (Supplementary Fig. 1 and Supplementary Videos 1–3)[12]. We therefore propose that this divalent cation could directly change the local electrostatic environment and interfere with the ionic $CD3_{CD}$–membrane interaction, thus facilitating ITAM phosphorylation.

We first performed a microdialysis assay to study whether Ca^{2+} could interfere directly with the $CD3_{CD}$–membrane interaction. The addition of Ca^{2+} significantly impaired the specific interaction of $CD3\varepsilon_{CD}$ or $CD3\zeta_{CD}$ to anionic phospholipids, even when the Ca^{2+} to lipid concentration ratio is only 0.1:1 (Fig. 1a, b and Supplementary Fig. 2). Another fluorescence polarization experiment also confirmed this observation (Supplementary Fig. 3). The effect of Ca^{2+} was still profound at physiological Ca^{2+} concentrations (Supplementary Fig. 4). We then performed a fluorescence resonance energy transfer (FRET)-based experiment to measure the effect of Ca^{2+} in live T cells. As described previously[1], the cytoplasmic domain of a monomeric membrane protein, KIR2DL3, was replaced by $CD3\varepsilon_{CD}$ with a carboxy-terminal teal fluorescence protein (monomeric TFP1). The distance between $CD3\varepsilon_{CD}$ and the plasma membrane was determined by measuring the FRET efficiency between TFP1 and a membrane dye, octadecyl rhodamine B (R18). To help translate the FRET efficiency

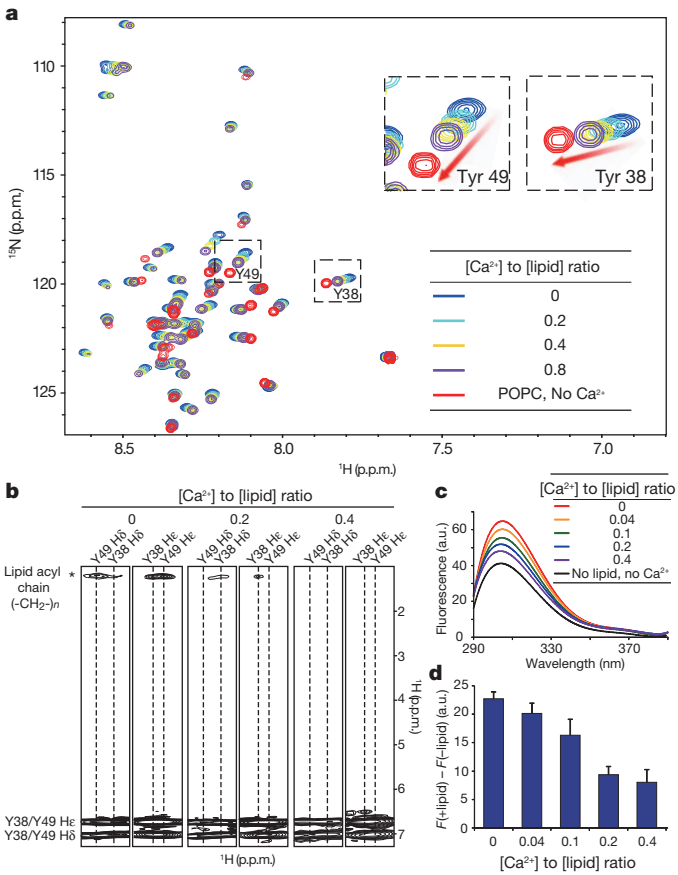


Figure 2 | Ca^{2+} induced the solvent exposure of tyrosine residues in $CD3\varepsilon_{CD}$ ITAM. **a**, Superimposed ^{15}N–^{1}H HSQC spectra of $CD3\varepsilon_{CD}$ with POPG bicelles (blue), $CD3\varepsilon_{CD}$ with POPG bicelles and Ca^{2+} with different concentrations (cyan, yellow and purple), and $CD3\varepsilon_{CD}$ with control POPC bicelles (red). Membrane-bound $CD3\varepsilon_{CD}$ (blue) and solvent-exposed $CD3\varepsilon_{CD}$ (red) had very different amide resonance spectra. Ca^{2+} was titrated into the $CD3\varepsilon_{CD}$-with-POPG sample at a molar ratio of $[Ca^{2+}]$:[POPG] from 0.2 to 0.8. In response to the increase in the Ca^{2+} concentration, the $CD3\varepsilon_{CD}$ amide resonances exhibited systematic shifts from the membrane-bound state to the solvent-exposed state. The resonance changes of two tyrosine residues, Tyr 38 and Tyr 49, are enlarged in the inset. HSQC experiments were performed with 0.05 mM ^{15}N-labelled $CD3\varepsilon_{CD}$ and 7.5 mM POPG or POPC bicelles ($q = 0.8$, q is the molar ratio of long-chain lipids (POPG or POPC) to short-chain lipids (DHPC)) in 20 mM Tris-HCl buffer, pH 7.0. **b**, Strips from aromatic NOESY spectra showing NOEs (distance of <5 Å) between the aromatic protons Hε/Hδ of Y38 and Y49 and the methylene protons in the lipid acyl chains. The substantial NOE signals (marked by asterisk) observed in the absence of Ca^{2+} indicated the insertion of tyrosine side chains into the membrane hydrophobic interior. The addition of Ca^{2+} resulted in the significant impairment of NOE signals, which indicated that tyrosine side chains were no longer embedded in the membrane core region. NOESY experiments were performed with 0.05 mM ^{15}N-, ^{13}C-labelled $CD3\varepsilon_{CD}$, 7.5 mM POPG bicelles ($q = 0.8$) and 0–3 mM Ca^{2+} in 20 mM Tris-HCl buffer, pH 7.0. Ca^{2+} was titrated in at a molar ratio relative to POPG of 0.2 to 0.4. **c**, Superimposed tyrosine fluorescence emission (TFE) spectra of $CD3\varepsilon_{CD}$ under different conditions. The peak value of the membrane-bound $CD3\varepsilon_{CD}$ TFE spectrum (red) was much higher than that of solvent-exposed $CD3\varepsilon_{CD}$ spectrum (black). Titrating in Ca^{2+} to the membrane-bound $CD3\varepsilon_{CD}$ sample substantially reduced the peak value (orange, green, blue and purple), confirming that Ca^{2+} induced the solvent exposure of tyrosine aromatic rings. **d**, The experiments shown in **c** were repeated three times, and the increase in the TFE value (at 310 nm) upon lipid binding, calculated as $F(+\text{lipid}) - F(-\text{lipid})$, was plotted as mean ± s.d.

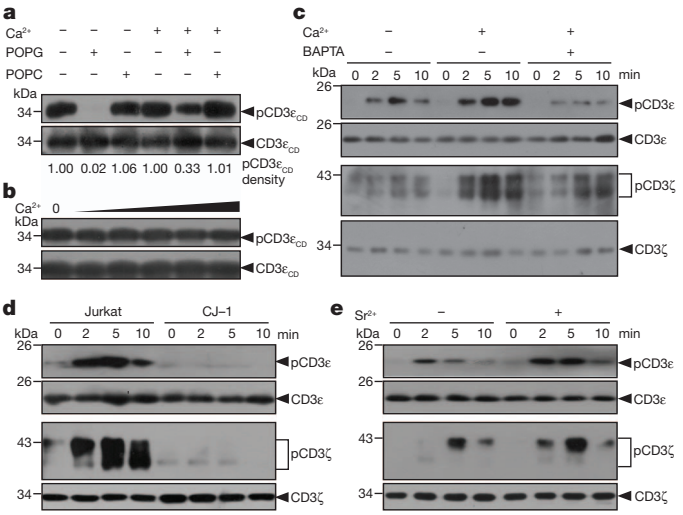


Figure 3 | Ca^{2+} facilitated CD3 phosphorylation. **a**, **b**, An *in vitro* phosphorylation assay was performed to study $CD3\varepsilon_{CD}$ tyrosine phosphorylation by Lck under different Ca^{2+} concentrations. **a**, The tyrosine residues of $CD3\varepsilon_{CD}$ were efficiently phosphorylated by Lck. The addition of POPG bicelles but not control POPC bicelles prevented $CD3\varepsilon_{CD}$ phosphorylation. Ca^{2+} can reverse this blockade and rescue $CD3\varepsilon_{CD}$ tyrosine phosphorylation. $pCD3\varepsilon_{CD}$, phosphorylated $CD3\varepsilon_{CD}$. **b**, Ca^{2+} did not directly increase the activity of Lck for phosphorylating $CD3\varepsilon_{CD}$ tyrosine residues. The Ca^{2+} concentration was raised from 0 to 0.5, 5, 50, 500 and 5,000 μM (left to right). **c**, Jurkat T cells were stimulated by TCR crosslinking (0.2 μg ml^{-1} anti-CD3 (UCHT1) plus 0.2 μg ml^{-1} anti-mouse IgG) for the indicated time at 37 °C. The levels of CD3ε and ζ phosphorylation were substantially greater in Ca^{2+}-containing triggering buffer (1 mM Ca^{2+}) compared with Ca^{2+}-free buffer. When intracellular Ca^{2+} was chelated with 10 μM BAPTA, the levels of both CD3ε and ζ phosphorylations were significantly reduced. **d**, CRAC-channel-deficient Jurkat mutant CJ-1 cells only had weak CD3ε and ζ phosphorylation after TCR crosslinking, indicating that Ca^{2+} influx was important for CD3 phosphorylation. **e**, To essentially rule out the potential contribution of other Ca^{2+}-dependent signalling pathways in CD3 phosphorylation, non-physiological Sr^{2+} (0.5 mM) was used instead of Ca^{2+} (1 mM) as the only divalent cation in the triggering buffer. A stable Jurkat cell line overexpressing constitutively active Lck (Supplementary Fig. 10b) was used for this experiment. Cells were stimulated using the same condition as in **c**. The levels of CD3ε and ζ phosphorylation were substantially greater under Sr^{2+}-containing condition compared with Sr^{2+}-free condition, even when Lck activity was already saturated.

into spatial distance, two control constructs with flexible linkers of different lengths (3 and 50 amino acids (aa), respectively) between KIR2DL3 transmembrane domain and TFP1 were used. The FRET efficiencies of these two constructs represent the states in which $CD3\varepsilon_{CD}$ is fully membrane-bound (3 aa) or unbound (50 aa). In resting cells, $CD3\varepsilon_{CD}$–TFP1 showed substantial FRET efficiency, similar to that of 3 aa–TFP1, indicating that $CD3\varepsilon_{CD}$ bound to the plasma membrane. After Ca^{2+} influx induced by ionomycin treatment or TCR crosslinking, the FRET efficiency of $CD3\varepsilon_{CD}$-TFP1 declined significantly to the level similar to that of 50 aa–TFP1 (Fig. 1e, f and Supplementary Fig. 5). In contrast, the FRET efficiencies of both 3 aa–TFP1 and 50 aa–TFP1 were not sensitive to ionomycin treatment (Fig. 1c, d, f). These data support the notion that Ca^{2+} influx can disrupt the ionic $CD3\varepsilon_{CD}$–lipid interaction and result in the dissociation of $CD3\varepsilon_{CD}$ from the plasma membrane.

We further used solution-state multi-dimensional NMR spectroscopy to trace the dynamics of the $CD3\varepsilon_{CD}$–membrane interaction in response to increase in Ca^{2+} concentration. As measured by a ^{15}N–^{1}H heteronuclear single-quantum coherence (HSQC) experiment, titrating in Ca^{2+} caused systematic shifts of amide resonances of all $CD3\varepsilon_{CD}$ residues, including the key ITAM tyrosines, Tyr 38 and Tyr 49, from membrane-bound state to solution state (Fig. 2a and Supplementary Fig. 6a). $CD3\zeta_{CD}$ showed similar sensitivity to increase in Ca^{2+} concentration (Supplementary Fig. 7). As one of the major groups interacting with anionic lipids, the side-chain amine groups of most arginines in $CD3\varepsilon_{CD}$ showed intense peaks in the membrane-bound $CD3\varepsilon_{CD}$ HSQC spectrum. Those peaks were largely missing in the soluble $CD3\varepsilon_{CD}$ spectrum because of the fast proton exchange with the solvent. The addition of Ca^{2+} to the membrane-bound $CD3\varepsilon_{CD}$ sample greatly reduced the signal intensity of the arginine side-chain amine groups, again confirming the ability of Ca^{2+} to induce the solvent exposure of $CD3\varepsilon_{CD}$ (Supplementary Fig. 6b). This effect was not caused by a Ca^{2+}-induced $CD3\varepsilon_{CD}$ conformational change because adding Ca^{2+} to the $CD3\varepsilon_{CD}$ solution in the absence of bicelles did not result in any specific chemical shift perturbation (Supplementary Fig. 6c). Our circular dichroism experiment also verified the effect of Ca^{2+} on disrupting the ionic $CD3\varepsilon_{CD}$–membrane interaction (Supplementary Fig. 8). Upon membrane binding, the aromatic rings of tyrosines insert into the hydrophobic core of the lipid bilayer, and substantial nuclear Overhauser effect (NOE) signals (distance <5 Å) between the tyrosine aromatic protons and the lipid acyl chain protons were detected (Fig. 2b). Adding Ca^{2+} substantially weakened the tyrosine–lipid NOE signals, indicating that the tyrosine side chains were no longer embedded in the membrane core region. A tyrosine fluorescence emission experiment confirmed the Ca^{2+}-induced solvent exposure of ITAM tyrosines (Fig. 2c, d).

We then tested whether Ca^{2+} could facilitate CD3 phosphorylation by both *in vitro* phosphorylation (IVP) and live T-cell stimulation experiments. The IVP data showed that binding to anionic lipids prevented $CD3\varepsilon_{CD}$ phosphorylation by Lck, but Ca^{2+} could reverse this blockade and rescue $CD3\varepsilon_{CD}$ phosphorylation (Fig. 3a). This is not owing to the enhancement of Lck activity by an increase in Ca^{2+} concentration (Fig. 3b). The physiological foreign antigen density on the antigen-presenting cell surface is generally very low[17]. To mimic

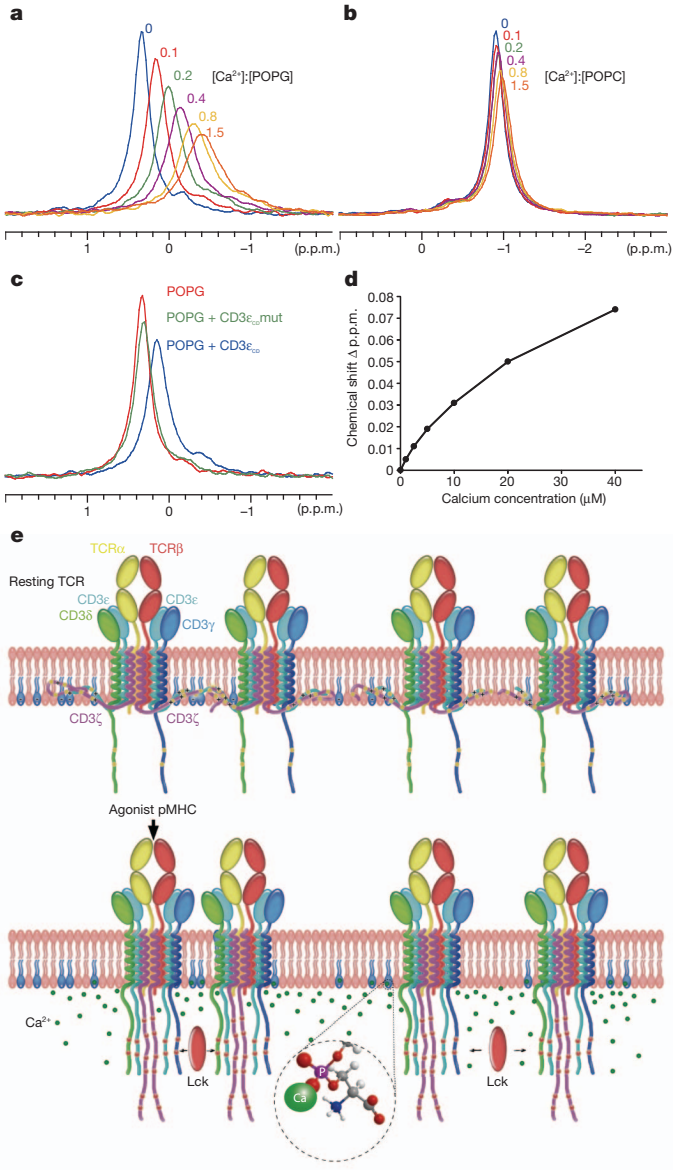


Figure 4 | Ca^{2+} bound to the phosphate group of anionic phospholipids at physiological concentrations. **a**–**d**, One-dimensional ^{31}P NMR experiments were performed to probe the local chemical environment change of the phospholipid phosphate group induced by the binding of Ca^{2+} or $CD3\varepsilon_{CD}$ peptide. A lipid nanodisc system was used to provide a membrane bilayer environment without the interference of detergent molecules. The sample buffer was 20 mM Tris-HCl, pH 7.0, containing 100 mM NaCl. **a**, **b**, Ca^{2+} induced a substantial chemical shift change of the phosphorus signal of the anionic POPG (**a**), but not that of the zwitterionic POPC (**b**). **c**, The $CD3\varepsilon_{CD}$ peptide induced a shift in the same direction for the POPG phosphorus signal as Ca^{2+} did, implying that both Ca^{2+} and basic residues in $CD3\varepsilon_{CD}$ bound to the same region in the phospholipid headgroup via ionic interactions. Mutation of the first three basic residues to Ala in $CD3\varepsilon_{CD}$ ($CD3\varepsilon_{CD}$ mutant) nearly eliminated the perturbation effect of $CD3\varepsilon_{CD}$ to the lipid phosphorus signal. **d**, The effect of Ca^{2+} in perturbing the ^{31}P signal was tested at physiological Ca^{2+} concentrations. Even 1 μM Ca^{2+} can induce unambiguous chemical shift changes of the POPG phosphorus signal. The POPG concentration used in **a** and **c** was 4 mM; in **d** the concentration was 0.8 mM. The POPC concentration used in **b** was 5 mM. The $CD3\varepsilon_{CD}$ concentration used in **c** was 0.08 mM. **e**, A schematic illustration of the Ca^{2+}-induced TCR signalling amplification model. In the resting state, the positively charged CD3ε/ζ cytoplasmic domains interact with anionic phospholipids in the inner leaflet of the plasma membrane and key tyrosine residues are sequestered in the membrane bilayer, which provides a 'safety' control on TCR triggering[1,8–10]. It has been well studied that Ca^{2+} influx starts within a few seconds after the initial TCR triggering[22] and the influx persists for several hours in synapsed T cells[29]. Membrane-proximal Ca^{2+} ions can then bind to the phosphate group in anionic phospholipids and neutralize their negative charges, which results in the dissociation of CD3ε/ζ cytoplasmic domains from the membrane and increases the accessibility of ITAM for Lck. Ca^{2+} can thus facilitate the phosphorylation of TCR–CD3 complexes, especially those contacting with low-affinity self antigens or even bystanders. This feedback regulation of Ca^{2+} can amplify the initial antigen-stimulated signal to a greater magnitude, which helps explain the unique nature of the hypersensitivity of T cells to even a single antigen. Moreover, the persistent high Ca^{2+} concentration in synapsed T cells could help sustain TCR signalling at a certain level, which is required for the full effector potential of T cells[29].

physiological conditions, we activated T cells with anti-CD3 antibody at a low concentration. For both Jurkat and mouse primary T cells, CD3 phosphorylation was substantially stronger when Ca^{2+} was present in the stimulating buffer (Fig. 3c and Supplementary Fig. 9a). Chelating intracellular Ca^{2+} with BAPTA (1,2-bis(o-aminophenoxy) ethane-N,N,N′,N′-tetraacetic acid) significantly reduced CD3 phosphorylation. CJ-1, a Jurkat mutant cell line having only residual expression of CRAC channel[23,24], had much weaker CD3 phosphorylation after T-cell activation when compared with wild-type Jurkat cells (Fig. 3d). Moreover, the presence of Ca^{2+} in the stimulating buffer did not cause substantial enhancement of CD3 phosphorylation in CJ-1 cells (Supplementary Fig. 9b). The effect of Ca^{2+} on CD3 phosphorylation is not caused by the direct enhancement of Lck activity or the suppression of tyrosine phosphatase activity. When active Lck was saturated or major tyrosine phosphatases were inhibited in T cells, the effect of Ca^{2+} on CD3 phosphorylation was still profound (Supplementary Figs 10 and 11). To further demonstrate that the role of Ca^{2+} in CD3 phosphorylation is primarily contributed by its charge, not other Ca^{2+}-dependent signalling pathways, we replaced Ca^{2+} with Sr^{2+}, a non-physiological divalent cation that can also pass through the CRAC channel[25] but is unable to trigger Ca^{2+}-dependent signalling pathways in T cells (Supplementary Fig. 12). Similar to Ca^{2+}, Sr^{2+} disrupted the $CD3\varepsilon_{CD}$–membrane interaction and facilitated CD3 phosphorylation (Fig. 3e, Supplementary Fig. 13).

We next studied the mechanism by which Ca^{2+} undermines the ionic $CD3\varepsilon_{CD}$–lipid interaction. We used ^{31}P NMR spectroscopy to probe directly the change in the chemical environment of the negatively charged lipid phosphorus after Ca^{2+} binding. Lipid nanodiscs were used here instead of bicelles to avoid the signal interference from the detergent molecule[26]. Ca^{2+}-binding induced substantial chemical shift change of the phosphorus signal of anionic lipids, but not that of the zwitterionic POPC (1-palmitoyl-2-oleoyl-*sn*-glycero-3-phosphocholine; Fig. 4a, b and Supplementary Fig. 14a, b). This perturbation effect can be reversed by the Ca^{2+} chelator EGTA (ethylene glycol tetraacetic acid; Supplementary Fig. 14c). Same directional shift of the phosphorus signal was induced by $CD3\varepsilon_{CD}$ binding, indicating that both Ca^{2+} and basic residues in $CD3\varepsilon_{CD}$ bound to the same lipid phosphate region (Fig. 4c). The Ca^{2+}–lipid phosphate binding was further confirmed by a Fourier transform infrared experiment (Supplementary Fig. 15). We then tested the effect of Ca^{2+} at low micromolar level to mimic the physiological Ca^{2+} concentration in activated T cells[15]. An unambiguous perturbation effect on the ^{31}P signal was observed even at 1 μM Ca^{2+}, indicating that the binding of Ca^{2+} and the lipid phosphate group is physiologically relevant (Fig. 4d). This binding thus explains a mechanism responsible for the role of Ca^{2+} in disrupting the ionic $CD3_{CD}$–membrane interaction and facilitating CD3 phosphorylation. As expected, Sr^{2+} can also specifically perturb the ^{31}P signal of anionic lipids (Supplementary Fig. 16).

This study unveils a new function of Ca^{2+} for feedback amplifying initial antigen-stimulated TCR signalling (Fig. 4e, Supplementary Discussion and Supplementary Fig. 17). This feedback loop will not cycle infinitely because the local Ca^{2+} concentration is precisely controlled by the plasma membrane Ca^{2+} ATPase pumps, mitochondria and other ion channels[15,27]. Moreover, TCRs become internalised and degraded when they move to the central immunological synapse[28]. The Ca^{2+}-mediated signal amplification process helps explain the unique nature of the hypersensitivity of T cells to even a single antigen molecule[11,18]. Ca^{2+} should also help sustain TCR signalling because the elevation of Ca^{2+} concentration persists for several hours in synapsed T cells[15,29]. Future work will be needed to fully demonstrate the *in vivo* relevance of our model and its potential applications to other signalling pathways involving ionic protein–lipid interactions[2–7,30].

METHODS SUMMARY

Mouse primary $CD4^+$ T cells were sorted from C57/B6 mice splenocytes using $CD4^+$ T cell Microbeads from Miltenyi Biotec. A CRAC-deficient Jurkat cell line CJ-1 was generously provide by R. Lewis. Jurkat and mouse $CD4^+$ primary T cells were stimulated by anti-CD3 crosslinking (for Jurkat cells, 0.2 μg ml^{-1} UCHT1 plus 0.2 μg ml^{-1} anti-mouse IgG; for primary T cells, 0.5 μg ml^{-1} 145-2C11 plus 1 μg ml^{-1} anti-hamster IgG) for the indicated time in Mg^{2+}-free Ringer's buffer with or without 1 mM Ca^{2+}. To chelate intracellular Ca^{2+}, T cells were pretreated with 10 μM BAPTA-AM (Invitrogen) for 30 min at 37 °C before antibody stimulation. CD3ε was immunoprecipitated with UCHT1 or 145-2C11 and analysed to determine its tyrosine phosphorylation status by anti-pY100 immunoblotting (Cell Signaling Technology). The tyrosine phosphorylation of CD3ζ was detected by anti-pCD3ζ (Santa Cruz Biotechnology) immunoblotting of cell lysate samples.

All NMR spectra were acquired at 27 °C on Bruker Avance 600 MHz and 850 MHz spectrometers. The acquired data were further processed using the software package NMRPipe and analysed with the program Sparky (Goddard & Kneller, University of California, San Francisco, California, USA). Due to the low concentration of protein/lipid/Ca^{2+} mixture samples, two-dimensional proton NOESY experiments were conducted instead of three-dimensional ^{13}C-separated NOESY experiments to observe the NOEs from $CD3\varepsilon_{CD}$ aromatic protons to lipid methyl protons. The binding of Ca^{2+} to the phosphate group of the phospholipid was probed by a one-dimensional ^{31}P NMR measurement on the 850 MHz spectrometer equipped with a ^{31}P direct-detection channel. Nanodiscs composed of different phospholipids (1-palmitoyl-2-oleoyl-*sn*-glycero-3-phospho-(1'-*rac*-glycerol) (POPG), 1-palmitoyl 2-oleoyl-*sn*-glycero-3-phospho-L-serine (POPS) or POPC) were used for the ^{31}P experiment to eliminate the interference of the phosphorus signal from the detergent molecule that is commonly used in other lipid bilayer systems[26]. ^{31}P spectra were measured with 256 accumulated transients.

Full Methods and any associated references are available in the online version of the paper.

Received 17 February; accepted 24 October 2012.
Published online 2 December 2012.

1. Xu, C. *et al.* Regulation of T cell receptor activation by dynamic membrane binding of the CD3epsilon cytoplasmic tyrosine-based motif. *Cell* **135,** 702–713 (2008).
2. Paddock, C. *et al.* Residues within a lipid-associated segment of the PECAM-1 cytoplasmic domain are susceptible to inducible, sequential phosphorylation. *Blood* **117,** 6012–6023 (2011).
3. Hansen, S. B., Tao, X. & MacKinnon, R. Structural basis of PIP2 activation of the classical inward rectifier K^+ channel Kir2.2. *Nature* **477,** 495–498 (2011).
4. Whorton, M. R. & MacKinnon, R. Crystal structure of the mammalian GIRK2 K^+ channel and gating regulation by G proteins, PIP2, and sodium. *Cell* **147,** 199–208 (2011).
5. Kim, C. *et al.* Basic amino-acid side chains regulate transmembrane integrin signalling. *Nature* **481,** 209–213 (2012).
6. van den Bogaart, G. *et al.* Membrane protein sequestering by ionic protein–lipid interactions. *Nature* **479,** 552–555 (2011).
7. Heo, W. D. *et al.* PI(3,4,5)P3 and PI(4,5)P2 lipids target proteins with polybasic clusters to the plasma membrane. *Science* **314,** 1458–1461 (2006).
8. Kuhns, M. S. & Davis, M. M. The safety on the TCR trigger. *Cell* **135,** 594–596 (2008).
9. DeFord-Watts, L. M. *et al.* The CD3ζ subunit contains a phosphoinositide-binding motif that is required for the stable accumulation of TCR–CD3 complex at the immunological synapse. *J. Immunol.* **186,** 6839–6847 (2011).
10. Aivazian, D. & Stern, L. J. Phosphorylation of T cell receptor ζ is regulated by a lipid dependent folding transition. *Nature Struct. Biol.* **7,** 1023–1026 (2000).
11. Irvine, D. J., Purbhoo, M. A., Krogsgaard, M. & Davis, M. M. Direct observation of ligand recognition by T cells. *Nature* **419,** 845–849 (2002).
12. Lioudyno, M. I. *et al.* Orai1 and STIM1 move to the immunological synapse and are up-regulated during T cell activation. *Proc. Natl Acad. Sci. USA* **105,** 2011–2016 (2008).
13. Smith-Garvin, J. E., Koretzky, G. A. & Jordan, M. S. T cell activation. *Annu. Rev. Immunol.* **27,** 591–619 (2009).
14. Palacios, E. H. & Weiss, A. Function of the Src-family kinases, Lck and Fyn, in T-cell development and activation. *Oncogene* **23,** 7990–8000 (2004).
15. Hogan, P. G., Lewis, R. S. & Rao, A. Molecular basis of calcium signaling in lymphocytes: STIM and ORAI. *Annu. Rev. Immunol.* **28,** 491–533 (2010).
16. Weiss, A. & Littman, D. R. Signal transduction by lymphocyte antigen receptors. *Cell* **76,** 263–274 (1994).
17. van der Merwe, P. A. & Dushek, O. Mechanisms for T cell receptor triggering. *Nature Rev. Immunol.* **11,** 47–55 (2011).
18. Purbhoo, M. A., Irvine, D. J., Huppa, J. B. & Davis, M. M. T cell killing does not require the formation of a stable mature immunological synapse. *Nature Immunol.* **5,** 524–530 (2004).
19. Sigalov, A. B., Aivazian, D. A., Uversky, V. N. & Stern, L. J. Lipid-binding activity of intrinsically unstructured cytoplasmic domains of multichain immune recognition receptor signaling subunits. *Biochemistry* **45,** 15731–15739 (2006).
20. Leventis, P. A. & Grinstein, S. The distribution and function of phosphatidylserine in cellular membranes. *Annu. Rev. Biophys.* **39,** 407–427 (2010).

21. Nika, K. *et al.* Constitutively active Lck kinase in T cells drives antigen receptor signal transduction. *Immunity* **32,** 766–777 (2010).
22. Huse, M. *et al.* Spatial and temporal dynamics of T cell receptor signaling with a photoactivatable agonist. *Immunity* **27,** 76–88 (2007).
23. Fanger, C. M., Hoth, M., Crabtree, G. R. & Lewis, R. S. Characterization of T cell mutants with defects in capacitative calcium entry: genetic evidence for the physiological roles of CRAC channels. *J. Cell Biol.* **131,** 655–667 (1995).
24. Park, C. Y., Shcheglovitov, A. & Dolmetsch, R. The CRAC channel activator STIM1 binds and inhibits L-type voltage-gated calcium channels. *Science* **330,** 101–105 (2010).
25. Yeromin, A. V. *et al.* Molecular identification of the CRAC channel by altered ion selectivity in a mutant of Orai. *Nature* **443,** 226–229 (2006).
26. Denisov, I. G., Grinkova, Y. V., Lazarides, A. A. & Sligar, S. G. Directed self-assembly of monodisperse phospholipid bilayer nanodiscs with controlled size. *J. Am. Chem. Soc.* **126,** 3477–3487 (2004).
27. Quintana, A. *et al.* Calcium microdomains at the immunological synapse: how ORAI channels, mitochondria and calcium pumps generate local calcium signals for efficient T-cell activation. *EMBO J.* **30,** 3895–3912 (2011).
28. Varma, R., Campi, G., Yokosuka, T., Saito, T. & Dustin, M. L. T cell receptor-proximal signals are sustained in peripheral microclusters and terminated in the central supramolecular activation cluster. *Immunity* **25,** 117–127 (2006).
29. Huppa, J. B., Gleimer, M., Sumen, C. & Davis, M. M. Continuous T cell receptor signaling required for synapse maintenance and full effector potential. *Nature Immunol.* **4,** 749–755 (2003).
30. Zilly, F. E. *et al.* Ca^{2+} induces clustering of membrane proteins in the plasma membrane via electrostatic interactions. *EMBO J.* **30,** 1209–1220 (2011).

Supplementary Information is available in the online version of the paper.

Acknowledgements We would like to thank R. Lewis for his gift of CJ-1 Jurkat cell line; S.-c. Sun, A. Lin, D. Li, J. J. Chou, H. Gu and M. Lei for discussions. C.X. is funded by the National Basic Research Program of China (973 Program, no. 2011CB910901 and no. 2012CB910804), the National Science Foundation of China (no. 31070738), the Chinese Academy of Sciences (Hundred Talents Program and no. KSCX2-EW-J-11), the Shanghai Municipal Commission for Science and Technology (10PJ1411500) and the Young Talent Program of Shanghai Institutes for Biological Sciences, CAS (no. 2010KIP101). J.W. is funded by the National Basic Research Program of China (973 Program, no. 2012CB917202) and the Chinese Academy of Sciences (Hundred Talents Program).

Author Contributions C.X. and J.W. conceived the project. X.S. and W.Y. performed the biochemical and T-cell activation experiments. Y.Bi and X.S. performed NMR, circular dichroism and Fourier transform infrared experiments. X.G. and Y.J. performed the FRET experiment. W.L. and C.X. supervised the TIRFM and spinning-disk confocal microscopy experiments. X.G. performed these imaging experiments and X.C. participated in the initial part of these imaging experiments. L.L., Y.Bai and J.G. helped on protein sample preparation. C.W., Y.W., B.W. and H.S. helped on nanodisc sample preparation. C.X. wrote the manuscript. J.W. and other authors revised the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.W. (junfeng@hmfl.ac.cn) or C.X. (cqxu@sibcb.ac.cn).

METHODS

Cells and reagents. Mouse $CD4^+$ T cells were sorted from C57/B6 mouse splenocytes using $CD4^+$ T cell selection microbeads (Miltenyi Biotec). A CRAC-deficient Jurkat cell line, CJ-1, was provided by R. Lewis. Lipids and detergent (POPG, POPC, POPS, DHPC (dihexanoyl-phosphatidylcholine)) were from Avanti polar lipids. Anti-CD3ε (M20), -phospho-CD3ζ, -CD3ζ (6B10.2) and -glutathione-*S*-transferase (GST) (Z-5) antibodies were from Santa Cruz Biotechnology. Phospho-Y100 antibody was from Cell Signaling Technology, UCHT1 from eBioscience, 145-2C11 from Biolegend, HA tag antibody (3F10) from Roche. Isotopes for NMR experiments were from Cambridge Isotope Laboratories.

Preparation of antibody-coated chambered coverglass slides. Lab-Tek chambered coverglass slides (Thermo Fisher Scientific) were cleaned by a strip buffer (H_2SO_4:H_2O_2 volume ratio 7:3), washed, and dried. 200 µl anti-human CD3 antibody UCHT1 (40 µg ml^{-1} in PBS) was loaded onto the coverglass slides and incubated at 37 °C for 30 min. After mild washing with PBS, the coverglass slides were blocked with 5% BSA in Ringer's buffer at 37 °C for 2 h and then mildly washed with Ringer's buffer before usage.

High-resolution high-speed time-lapse live cell imaging by total internal reflection fluorescence microscopy (TIRFM) and spinning-disk confocal microscopy. GCaMP5, a new generation of genetically encoded Ca^{2+} indicator (Looger Lab, Janelia Farm, HHMI), was used to image Ca^{2+} in live T cells. GCaMP5- or HA–human CD3ε–GCaMP5-expressing Jurkat transfectants were first loaded with 100 µM EGTA-AM (a membrane-permeable form of EGTA; Invitrogen) and resuspended in Ringer's Buffer. EGTA was not used for the spinning-disk confocal imaging of HA–hCD3ε–GCaMP5-expressing cells. Cells were placed on antibody-coated coverglass slides right before imaging. TIRF images were acquired at 37 °C on a heated stage by an Olympus IX-81 microscope supported by a TIRF port, ANDOR iXon+ DU897D electron-multiplying charge-coupled device (EMCCD) camera (ANDOR Technology) and Olympus ×100 1.45 numerical aperture (NA) lenses. The acquisition was controlled by MetaMorph software (Molecular Devices). The exposure time was 100 ms unless otherwise indicated. A Coherent Sapphire 488 nm diode laser was used. Spinning-disk confocal images were acquired by an Olympus IX-81 microscope supported by the Yokogawa spinning-disk sets, ANDOR iXon+ DU897D EMCCD and Olympus ×100 1.45 NA lenses. The acquisition was controlled by ANDOR Q2 software (ANDOR Technology). The exposure time and laser used were the same with above TIRF set. Images were analysed with Image Pro Plus software (Media Cybernetics).

$CD3\varepsilon_{CD}$ and $CD3\zeta_{CD}$ peptide expression and labelling. $CD3\varepsilon_{CD}$ and $CD3\zeta_{CD}$ were expressed as a GST fusion protein in *Escherichia coli*, separated from GST by tobacco etch virus protease cleavage and purified by reverse-phase HPLC. For microdialysis and fluorescence polarization experiments, peptides were expressed with an additional amino-terminal cysteine and labelled by an Alexa488-maleimide derivative (Invitrogen). For the *in vitro* phosphorylation experiment, the GST–$CD3\varepsilon_{CD}$ fusion protein was used directly for the unbiased detection of phosphorylated and unphosphorylated $CD3\varepsilon_{CD}$ by anti-GST immunoblotting.

Equilibrium-based microdialysis assay. The microdialysis system was composed of two equal size chambers, an experimental chamber and a control chamber[1]. The experimental chamber was loaded with 10 nM Alexa488-labelled $CD3\varepsilon_{CD}$ or $CD3\zeta_{CD}$ and large unilamellar vesicles (LUVs) composed of 5 mM POPG (an anionic phospholipid) in buffer containing 20 mM HEPES, pH 7.0 and 100 mM KCl, whereas the control chamber was loaded with 10 nM Alexa488-labelled $CD3\varepsilon_{CD}$ or $CD3\zeta_{CD}$ and LUVs composed of 5 mM POPC (a zwitterionic phospholipid) in the same buffer. Ca^{2+} was added into the two chambers at a molar ratio relative to the phospholipid from 0 to 1. The two chambers were separated by a 300-kDa dialysis membrane, which allowed the free movement of the fluorescent peptides but not the LUVs. After reaching equilibrium at 16 h, the binding of $CD3\varepsilon_{CD}$ or $CD3\zeta_{CD}$ to the anionic phospholipid was calculated on the basis of the fluorescence intensity difference between the two chambers. The binding efficiency of $CD3\varepsilon_{CD}$ or $CD3\zeta_{CD}$ to acidic phospholipids was calculated by the following equation: Binding (%) = (FLExp − FLCon)/(FLExp + FLCon) × 100, where FLExp is the fluorescence intensity in the experimental chamber and FLCon is the fluorescence intensity in the control chamber. The binding efficiencies under different Ca^{2+} concentrations were then normalized to the value measured in the absence of Ca^{2+}.

To study the effect of Ca^{2+} on $CD3\varepsilon_{CD}$–lipid binding at physiological micromolar Ca^{2+} concentrations (Supplementary Fig. 4), we optimized the system and used 20 µM POPG or POPC LUV, 2 nM Alexa488-$CD3\varepsilon_{CD}$ peptide and 0–20 µM Ca^{2+} in 20 mM HEPES buffer, pH 7.0, containing 150 mM KCl.

Fluorescence polarization assay. To measure the binding kinetics of $CD3\varepsilon_{CD}$ to phospholipids and the inhibition of this binding by Ca^{2+}, Alexa488-labelled $CD3\varepsilon_{CD}$, POPG/DHPC bicelles ($q = 0.8$) and Ca^{2+} were incubated in a 384-well plate for 15 min at 27 °C. The sample buffer was 20 mM HEPES, pH 7.0, 150 mM NaCl. Fluorescence polarization values were measured using a Tecan Infinite M1000Pro Microplate Reader.

Tyrosine fluorescence emission assay. The tyrosine fluorescence emission spectrum of $CD3\varepsilon_{CD}$ was detected by a Varian Cary Eclipse machine, with the excitation wavelength 275 nm and the emission wavelength 290–390 nm. Mixture bicelles (30% POPS, 10% POPG and 60% POPC) were used to mimic the physiological lipid composition in the inner leaflet of the plasma membrane. Tyrosine fluorescence emission experiments were performed with 6 µM $CD3\varepsilon_{CD}$, 0.1 mM DTT, 0 or 0.5 mM POPS/POPG/POPC mixture bicelles ($q = 0.8$) and 10 mM Tris-HCl, pH 7.4. Ca^{2+} was titrated at a molar ratio relative to the total lipid concentration of 0.04 to 0.4.

FRET measurement. One million cells were resuspended in 1 ml of Mg^{2+}-free Ringer's buffer (155 mM NaCl, 4.5 mM KCl, 10 mM D-glucose, 10 mM HEPES, pH 7.4) containing 2 mM $CaCl_2$. A 300-µl aliquot was pipetted onto the centre of 35 mm glass-bottom dish (Shengyou Biotechnology) and cells were allowed to adhere for 5 min at room temperature. To induce Ca^{2+} influx, 5 µM ionomycin (Sigma) or anti-CD3 antibody (2 µg ml^{-1} UCHT1 and 2 µg ml^{-1} anti-mouse IgG) was added into the dish, which was then incubated at room temperature for additional 5 min. To label the plasma membrane, 0.5 µl of 500 µg ml^{-1} octadecyl rhodamine B (R18) (Invitrogen) was added into the dish. The dish was then mounted onto a Leica TCS SP5 microscope for imaging. The FRET efficiency between TFP1 (donor) and R18 (acceptor) was measured by the dequenching method[1]. TFP1 was excited with the Argon 458 nm laser line and visualized using detection at 470–560 nm. R18 was excited with the Helium/Neon 561 nm laser line, visualized using detection at 590–650 nm. Unpaired two tailed Student's *t*-test was used for the statistical analysis of all FRET data.

***In vitro* phosphorylation assay.** $CD3\varepsilon_{CD}$ with an N-terminal GST tag was used in the *in vitro* phosphorylation assay, for the unbiased detection of phosphorylated and unphosphorylated $CD3\varepsilon_{CD}$ peptides by anti-GST immunoblotting. GST–$CD3\varepsilon_{CD}$ (1 µM) was first incubated with lipids (2 mM POPG or POPC bicelles) in the presence or absence of 10 mM Ca^{2+} at 37 °C for 30 min. The reaction volume was 20 µl and the buffer was 20 mM HEPES, pH 7.0, 150 mM KCl. Lck (Millipore) and kinase buffer were then added to the reaction and the final volume was 40 µl, containing 0.5 µM $CD3\varepsilon_{CD}$, 1 mM lipid bicelles, 2 µg ml^{-1} Lck, 0 or 5 mM Ca^{2+}, 70 mM HEPES, pH 7.0, 5 mM MgCl2, 3 µM Na3VO4, 1.25 mM DTT, 150 mM KCl and 0.2 mM ATP. The reaction was incubated at 30 °C for 1 h. Tyrosine-phosphorylated and total $CD3\varepsilon_{CD}$ were detected by anti-pY100 and anti-GST immunoblotting, respectively.

To test whether Ca^{2+} concentration increase could directly enhance Lck activity, Lck (2 µg ml^{-1}) was incubated with $CD3\varepsilon_{CD}$ (0.5 µM) and Ca^{2+} (0–5 mM) at 30 °C for 1 h to phosphorylate $CD3\varepsilon_{CD}$ tyrosine residues.

T-cell activation and immunoblotting. Jurkat cells and mouse $CD4^+$ primary T cells were stimulated at 37 °C by TCR crosslinking (for Jurkat: 0.2 µg ml^{-1} anti-CD3 (UCHT1) plus 0.2 µg ml^{-1} anti-mouse IgG; for primary T cells, 0.5 µg ml^{-1} anti-CD3 (145-2C11) plus 1 µg ml^{-1} anti-hamster IgG) for the indicated time in Mg^{2+}-free Ringer's buffer with or without 1 mM Ca^{2+}. To chelate intracellular Ca^{2+}, the cells were pretreated with 10 µM BAPTA-AM (Invitrogen) for 30 min at 37 °C. Stimulation reactions were terminated by adding ice-cold Mg^{2+}-free Ringer's buffer. For the detection of CD3ε phosphorylation, cells were lysed in an immunoprecipitation buffer (1% NP40, 50 mM Tris-HCl, pH 7.4, 155 mM NaCl, 2 mM EDTA, 2 mM Na_3VO_4, 20 mM NaF, 10 mM iodoacetamide, 1 mM PMSF and complete protease inhibitor cocktail (Sigma)) and then subjected to CD3ε immunoprecipitation by UCHT1 or 145-2C11. For the detection of CD3ζ phosphorylation, stimulated cells were directly lysed in SDS–PAGE sample buffer. Immunoprecipitated samples and cell lysate samples were separated by SDS–PAGE, transferred onto PVDF membranes (Millipore) and subjected to immunoblotting. CD3ε phosphorylation and CD3ζ phosphorylation were detected by anti-pY100 and anti-pCD3ζ immunoblotting, respectively. Membranes were then stripped off and re-blotted with anti-CD3ε or anti-CD3ζ (6B10.2) for the detection of total CD3ε and CD3ζ, respectively.

HA crosslinking. To study how receptor aggregation affects ITAM–membrane interaction, we generated a chimaera construct, HA–KIR–$CD\zeta_{CD}$-TFP1, which contains an N-terminal HA tag, KIR2DL3 extracellular and transmembrane domains, CD3ζ cytoplasmic domain and a C-terminal monomeric TFP1. Stable Jurkat cell lines expressing this construct were generated by lentiviral infection and FACS sorting. For FRET measurement, one million Jurkat transfectants were resuspended in 1 ml Ca^{2+} and Mg^{2+}-free Ringer's buffer. A 300 µl cell suspension aliquot was pipetted onto the centre of 35 mm Glass bottom dish and cells were allowed to adhere for 5 min at the room temperature. To crosslink the monomeric CD3ζ chimaera, 3 µl of 20 µg ml^{-1} rat anti-HA antibody and 20 µg ml^{-1} anti-rat IgG mixture was added into the dish and incubated at the room temperature for additional 5 min. FRET measurements were performed as described above. For the

chimaera phosphorylation experiment, Jurkat transfectants were stimulated at 37 °C by anti-HA crosslinking (5 μg ml^{-1} anti-HA (3F10, Roche) + 10 μg ml^{-1} goat anti-rat IgG) for indicated times in Ca^{2+}, Mg^{2+}-free Ringer's buffer. Cells were then lysed and the chimaera was pulled down by anti-HA antibody immunoprecipitation. Immunoprecipitated samples were separated by SDS–PAGE, transferred onto PVDF membrane and subject for immunoblotting. CD3ζ_{CD} phosphorylation was detected by anti-pCD3ζ immunoblotting (pY142, Anbo biotech). Membranes were then stripped off and re-blotted by anti-HA for the detection of total HA-KIR-CD3ζ_{CD}-TFP1.

Preparation of nanodiscs, bicelles and LUVs. To assemble nanodiscs, the MSP1D1/lipid/sodium cholate reaction mixture was incubated for 15 h at the appropriate temperature (POPC 4 °C, POPG/POPS 14 °C) in buffer containing 20 mM Tris-HCl, pH 7.2, 100 mM NaCl and 0.5 mM EDTA[26]. The molar ratio of MSP1D1:lipid:sodium cholate was 1:65:130. The nanodisc self-assembly process was initiated upon the removal of the sodium cholate by incubating the mixture with Bio-Beads SM-2 (0.8 g per ml reaction mixture, Bio-Rad) on a shaker for 8 h at room temperature After filtering off Bio-Beads with syringe filters, the nanodisc samples were purified on a Superdex 200 gel-filtration column (GE Healthcare) equilibrated in buffer containing 20 mM Tris-HCl, pH 7.0 and 100 mM NaCl. The peak fractions were collected and concentrated using an ultracentrifugation column with a 30,000 MW cutoff (Millipore). Large unilamellar vesicle and bicelle samples were prepared as previously described[1].

NMR experiments. All NMR spectra were acquired at 27 °C on Bruker Avance 600 MHz and 850 MHz spectrometers. The acquired data were further processed using the software package NMRPipe[31] and analysed with Sparky (Goddard & Kneller, University of California, San Francisco, California, USA).

Aromatic NOESY. Because of the low concentration of the protein/lipid/Ca^{2+} mixture samples, two-dimensional proton NOESY experiments were conducted to observe NOEs from CD3ε_{CD} aromatic protons to lipid methyl protons. For the indirect dimension, background signals from CD3ε_{CD} were suppressed by double ^{13}C filtering, and only lipid methyl protons evolved. For the direct dimension, the aromatic protons Hδ and Hε from tyrosines were selected by aromatic ^{13}C edition. The samples were diluted in D_2O to achieve better water suppression and reduce the spin diffusion effect. Experiments were conducted on the Bruker 600 MHz spectrometer at 27 °C with a NOE mixing time of 300 ms, 128 points in the indirect dimension, 2,048 points in the direct dimension and 1,024 accumulated transients.

^{31}P spectrum. The binding of Ca^{2+} to the phosphate group of the phospholipid was measured by ^{31}P spectrum on the 850 MHz spectrometer equipped with a ^{31}P direct-detection channel. Lipid nanodisc samples were prepared in buffer containing 20 mM Tris-HCl, pH 7.0, 100–150 mM NaCl. Ca^{2+} was titrated in with a molar ratio of Ca^{2+}:lipid from 0 to 1.5. One-dimensional ^{31}P spectra were measured with 256 accumulated transients and processed with 25 Hz of line broadening.

31. Delaglio, F. *et al.* NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J. Biomol. NMR* **6,** 277–293 (1995).

LETTER

doi:10.1038/nature11675

Coordinated control of replication and transcription by a SAPK protects genomic integrity

Alba Duch[1], Irene Felipe-Abrio[2], Sonia Barroso[2], Gilad Yaakov[1], María García-Rubio[2], Andrés Aguilera[2], Eulàlia de Nadal[1] & Francesc Posas[1]

Upon environmental changes or extracellular signals, cells are subjected to marked changes in gene expression [1,2]. Dealing with high levels of transcription during replication is critical to prevent collisions between the transcription and replication pathways and avoid recombination events[3–5]. In response to osmostress, hundreds of stress-responsive genes are rapidly induced by the stress-activated protein kinase (SAPK) Hog1 (ref. 6), even during S phase[7]. Here we show in *Saccharomyces cerevisae* that a single signalling molecule, Hog1, coordinates both replication and transcription upon osmostress. Hog1 interacts with and phosphorylates Mrc1, a component of the replication complex[8–11]. Phosphorylation occurs at different sites to those targeted by Mec1 upon DNA damage[8,9]. Mrc1 phosphorylation by Hog1 delays early and late origin firing by preventing Cdc45 loading, as well as slowing down replication-complex progression. Regulation of Mrc1 by Hog1 is completely independent of Mec1 and Rad53. Cells carrying a non-phosphorylatable allele of *MRC1* (*mrc1^{3A}*) do not delay replication upon stress and show a marked increase in transcription-associated recombination, genomic instability and Rad52 foci. In contrast, *mrc1^{3A}* induces Rad53 and survival in the presence of hydroxyurea or methyl methanesulphonate. Therefore, Hog1 and Mrc1 define a novel S-phase checkpoint independent of the DNA-damage checkpoint that permits eukaryotic cells to prevent conflicts between DNA replication and transcription, which would otherwise lead to genomic instability when both phenomena are temporally coincident.

High levels of transcription can induce genomic instability owing to conflicts between transcription and replication[3]. SAPKs are key elements in intracellular signalling networks that serve to respond and adapt to extracellular changes. Exposure of yeast to high osmolarity activates the p38-related SAPK Hog1, which reprograms the expression capacity of the cell[6,12–14] and delays cell-cycle progression[15–17], raising the question of how cells make massive changes in gene expression compatible with DNA replication.

To identify molecular targets for Hog1 in the replication complex (RC), a subset of 30 proteins involved in replication were subjected to an *in vitro* phosphorylation assay with active Hog1. Only Mrc1, which couples the DNA helicase and the DNA polymerase during replication[8–11], was phosphorylated by Hog1 (Fig. 1a). Mrc1 was also *in vivo* phosphorylated upon osmostress in a Hog1-dependent manner (Fig. 1b). Mrc1 has three MAPK consensus sites (T169, S215 and S229), and mutations of all three sites mostly abolished phosphorylation by Hog1 (Fig. 1a, b and Supplementary Fig. 2). These phosphorylation sites are different to those involved in the intra-S-phase DNA checkpoint[8,9]. Then, we tested whether endogenous tagged Hog1 and Mrc1 interacted. Notably, Mrc1 precipitated with Hog1 and vice versa (Fig. 1c and Supplementary Fig. 3a). This interaction was confirmed by a two-hybrid assay (Supplementary Fig. 3b, c). These results indicate that Mrc1 is a bona fide target of Hog1.

To assess the relevance of Mrc1 and its phosphorylation by Hog1 in the cell cycle, cells were synchronized at the onset of S phase by using a temperature-sensitive *CDC7* allele (*cdc7-4*) after release from pheromone, and then released with or without osmostress (NaCl or sorbitol). S-phase progression was delayed in osmostressed cells (Fig. 2a) but the delay was not observed in *mrc1^{3A}* cells (Fig. 2a). Similar results were obtained after release from α-factor or *cdc7-4* synchronization (Supplementary Fig. 4a, b). Hog1 can be activated by the Pbs2 MAPKK[18] (Pbs2^{DD}). Overexpression of Pbs2^{DD} also induced an S-phase delay in *cdc7-4* cells that was abolished in *mrc1^{3A}* cells (Supplementary Fig. 5a, b). Clb5 is degraded after cells exit S phase[19,20].

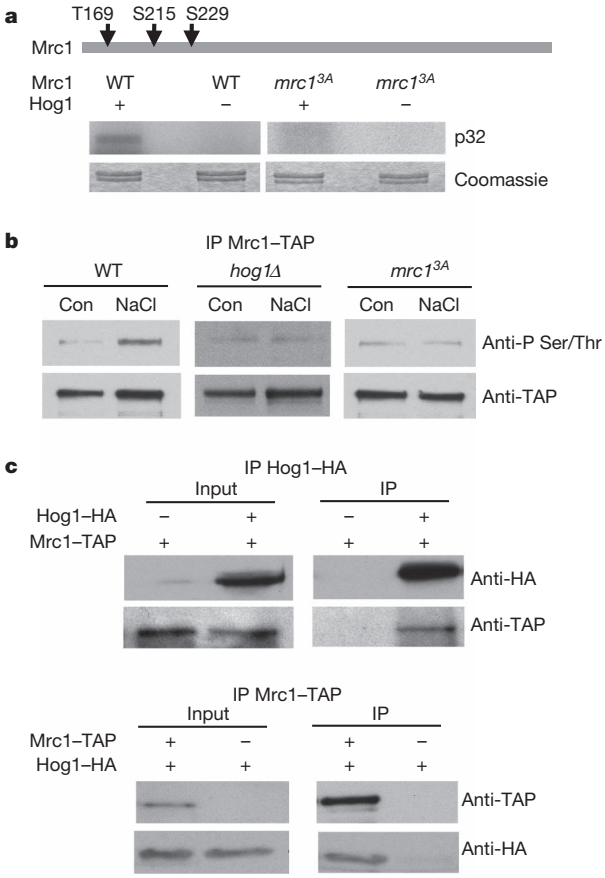


Figure 1 | Mrc1 is a Hog1 target. a, Hog1 phosphorylates Mrc1 *in vitro*. Purified glutathione *S*-transferase (GST)–Mrc1 protein is phosphorylated by Hog1 in an *in vitro* kinase assay, and this phosphorylation is abolished in GST–Mrc1^{3A}, which harbours mutations of Thr 169, Ser 215 and Ser 229 to Ala. WT, wild type. **b**, Mrc1 is a Hog1 target *in vivo*. Mrc1–TAP pulled down from S-phase-synchronized cultures is phosphorylated by Hog1 in osmostressed cells. Con, control; IP, immunoprecipitate. **c**, Hog1 and Mrc1 interact *in vivo*. Mrc1–TAP co-immunoprecipitates with Hog1–haemagglutinin (HA) and vice versa in *in vivo* pull-down assays.

[1]Cell Signaling Unit, Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra (UPF), Barcelona E-08003, Spain. [2]Centro Andaluz de Biología Molecular and Medicina Regenerativa CABIMER, Universidad de Sevilla, 41092 Sevilla, Spain.

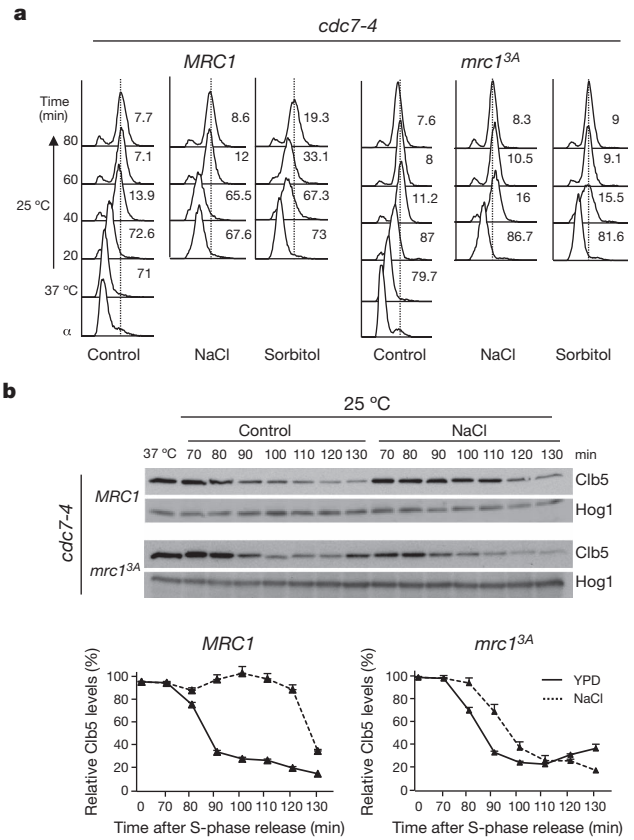


Figure 2 | *mrc1*^{3A} bypasses the osmostress-induced S phase delay. **a**, **b**, *cdc7-4* (*MRC1* or *mrc1*^{3A}) cells were pre-synchronized with α-factor (α), washed and shifted to 37 °C to synchronize them at the onset of S phase (37 °C). Cells were released at 25 °C in yeast extract peptone dextrose (YPD) (control), NaCl or sorbitol, and cell-cycle progression was analysed by fluorescence-activated cell sorting (FACS) (**a**) or by western blot (**b**). **a**, *mrc1*^{3A} cells cannot delay DNA replication upon osmostress. Vertical dotted line indicates the end of S phase (2C peak). The percentage of cells in S phase is shown at the top right of graphs. **b**, *mrc1*^{3A} cells show no delay in Clb5 degradation after release upon osmostress (NaCl). Bottom panels show normalized Clb5 quantification. Data represent the mean and standard deviation (s.d.) of three independent experiments.

Correspondingly, wild-type (*cdc7-4*) cells subjected to osmostress delayed Clb5 degradation in contrast to *mrc1*^{3A} cells (Fig. 2b). Of note, *mec1*- or *rad53*-deficient cells delayed S-phase progression upon osmostress as efficiently as wild-type cells (Supplementary Fig. 6a). In addition, Rad53 was not phosphorylated upon osmostress, in contrast to treatment with hydroxyurea (HU) (Supplementary Fig. 6b). This indicates that osmostress delays S phase independently of the S-phase checkpoint that responds to DNA damage. Correspondingly, cells containing the *mrc1*^{3A} allele blocked S-phase progression in response to HU, phosphorylated Rad53 and were resistant to HU and methyl methanesulphonate (MMS), showing an intact DNA-damage checkpoint (Supplementary Fig. 7). Therefore, phosphorylation of Mrc1 by Hog1 is essential for the S-phase delay upon osmostress but it is not required for the DNA-damage response.

To characterize the mechanism by which Hog1 phosphorylation of Mrc1 delays S phase, we followed the binding of endogenous tagged Dpb2 at replication origins and adjacent regions. The Dpb2 subunit of the DNA polymerase ε serves to monitor origin activation and RC progression[21]. Dpb2 binding at the early origin ARS305 was similar in control and stressed cells, but was clearly delayed in adjacent regions upon osmostress (8 kilobase pairs (kb) and 17 kb). This delay was barely present in *mrc1*^{3A} cells (Fig. 3a and Supplementary Fig. 8). Additionally, recruitment of Dpb2 to the late origin ARS501 was delayed in wild-type but not in mutant cells (Fig. 3a and Supplementary Fig. 8). Similar results were obtained when loading of Mrc1 or Pol2 were assessed (Supplementary Figs 9 and 10a).

We then analysed replication-fork progression by two-dimensional gel electrophoresis. Replication through the DNA region adjacent to ARS305 lasted longer in wild-type cells when compared to *mrc1*^{3A} cells, with the Y-arc persisting in the wild type while fading away in the mutant (Fig. 3b). Thus, Hog1 phosphorylation of Mrc1 delays fork progression upon stress. It has been suggested that the altered interaction between Mrc1 and the Pol2 subunit of DNA pol ε delays S-phase progression upon replication stress[22]. Binding of Mrc1 to Pol2 was reduced to less than 50% upon osmostress. In contrast, the association of Pol2 with the non-phosphorylatable mutant Mrc1^{3A} protein was not altered (Fig. 3c and Supplementary Fig. 11a). The association of Dpb2 with Mrc1 or Pol2 was also not altered (Supplementary Fig. 11b). Therefore, osmostress leads to reduced binding of Mrc1 to DNA Pol2, which might explain a decrease in RC progression.

The quantification of the two-dimensional gels also suggested that firing of replication at ARS305, as indicated by the appearance of the bubble arc, was slightly delayed in the presence of stress in wild-type in contrast to *mrc1*^{3A} cells. Then, we performed DNA-combing experiments using synchronous cultures to measure replication speed and RC progression (track length). Wild-type cells showed significantly shorter replication tracks upon osmostress (30 min) than *mrc1*^{3A} cells, suggesting that the latter was not able to delay early origin firing and RC progression (Fig. 3d). Similar results were observed in asynchronous cultures (Supplementary Fig. 12a). In addition, the assays performed at different time points (30 and 45 min) (Fig. 3d and Supplementary Fig. 12b) showed that wild-type cells showed reduced speed of replication upon osmostress when compared to *mrc1*^{3A} cells (0.3 kb min^{-1} versus 0.57 kb min^{-1}). Loading of the Cdc45 helicase is essential for origin firing. Remarkably, binding of Cdc45 to early origins was delayed in wild-type cells and in *rad53* cells, but not in *mrc1*^{3A} cells, upon osmostress (Fig. 3e and Supplementary Fig. 10c). Thus, Hog1 phosphorylation of Mrc1 delays origin firing by preventing the association of Cdc45 with early origins. In contrast, HU does not alter the association of Cdc45 with early origins, which is consistent with HU not preventing early origin firing (Supplementary Fig. 10b). Although it is possible that the phosphorylation of Mrc1 by Hog1 could lead to recruitment of additional factors to regulate origin firing, Mrc1 phosphorylation upon osmostress delays S-phase progression by altering early and late origin firing and possibly inhibiting fork progression.

We next asked how important it was for the cell to delay S-phase progression upon osmostress. To do this, we assayed chromosomal instability upon osmostress in wild-type or *mrc1*^{3A} cells by a sectoring colour assay (Methods). *mrc1*^{3A} cells showed a marked increase in chromosomal instability in response to osmostress that was not observed in wild type (Fig. 4a and Supplementary Fig. 13a). Correspondingly, an *MRC1* mutant carrying Hog1-phosphomimic amino acids (*mrc1*^{3D}) showed slower S-phase progression even in the absence of stress (Supplementary Fig. 14a), slower replication upon osmostress (Supplementary Fig. 14b) and did not have a significant increase in genomic instability upon osmostress, indicating that Mrc1 phosphorylation by Hog1 is essential to preserve genomic stability upon stress (Supplementary Fig. 13a).

High levels of transcription can induce genomic instability owing to collisions between transcription and replication complexes[23–25], and a major adaptive response to osmostress is the control of gene expression by Hog1 (ref. 6). It was therefore conceivable that concurrent transcription upon osmostress with ongoing replication could be provoking transcription-associated recombination (TAR)[23,26]. Thus, we asked whether TAR increased upon osmostress. A plasmid in which transcription of a *leu2* direct-repeat recombination system was driven by the *STL1* stress-responsive promoter in two different orientations (IN and OUT) with respect to the ARS209 (ARSH4) origin allowed

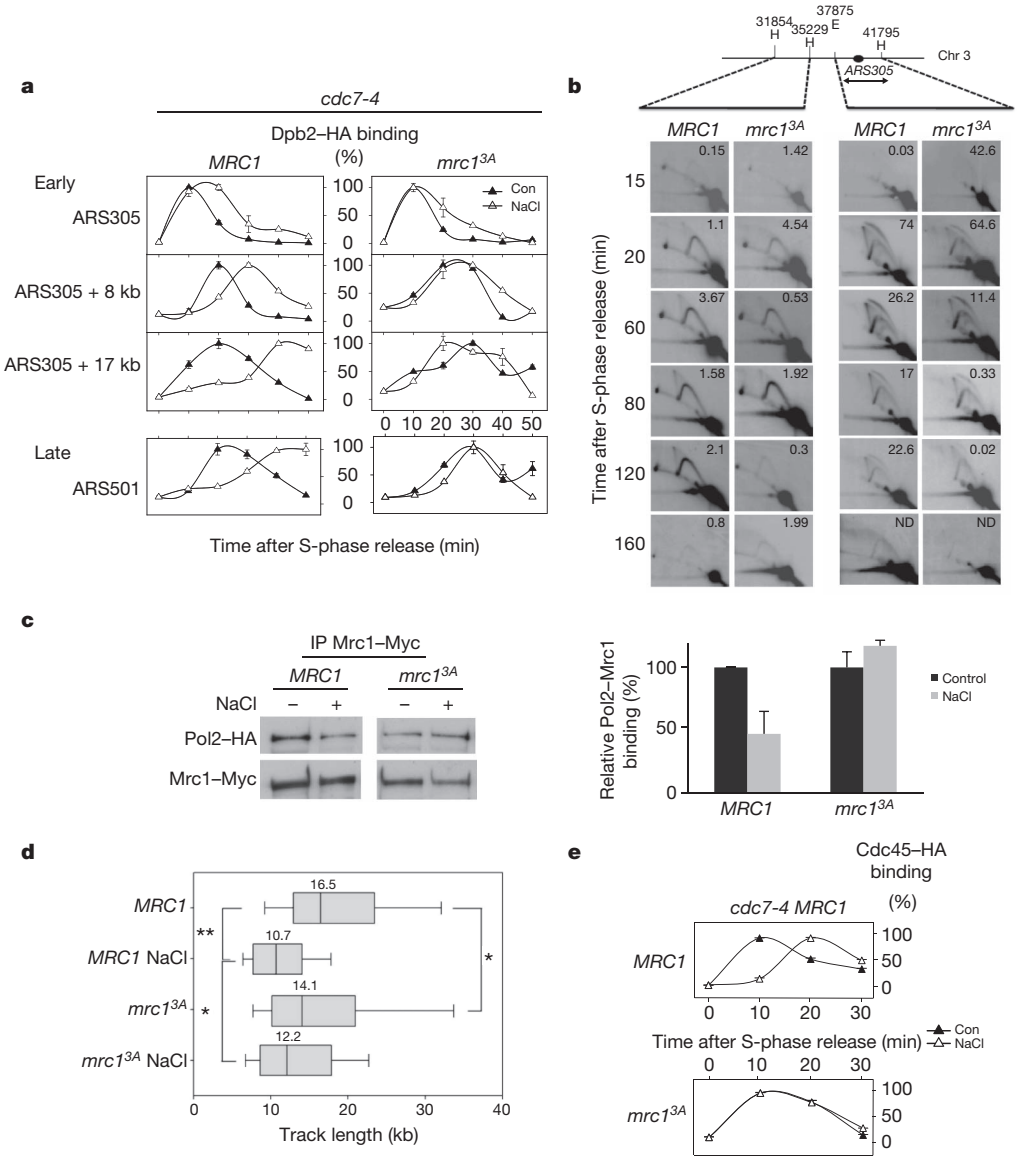


Figure 3 | Hog1 delays replication by inhibiting fork progression and origin firing. **a**–**e**, Wild-type (WT) and *mrc1*^{3A} cells were synchronized as in Fig. 2. **a**, Dpb2 binding at the indicated regions was determined by chromatin immunoprecipitation (ChIP). The graph shows the kinetics of Dpb2 after release in YPD (Con) or osmostress (NaCl). **b**, Two-dimensional gel electrophoresis analysis of replication-fork progression at early origin *ARS305* or at adjacent regions. Quantified data are indicated in top right corners as percentage of replication forks (adjacent region) or as percentage of bubbles among total replication forks (*ARS305*). ND, not determined. **c**, Co-immunoprecipitation experiments show that Hog1 phosphorylation of Mrc1 upon osmostress alters association with Pol2. IP, immunoprecipitate. **d**, DNA-combing analysis of replication forks in wild-type and *mrc1*^{3A} cells in the absence or presence of osmostress (30 min). Graph indicates the distribution of BrdU track length (kb). $*P < 0.05$, $**P < 0.005$. **e**, Hog1 phosphorylation of Mrc1 delays Cdc45 binding at the early origin *ARS305*. Data show the mean and s.d. of three independent experiments.

us to assess recombination of the *leu2* direct repeat (Fig. 4b, Supplementary Fig. 13b and Methods). In the absence of stress, neither the wild-type nor the *mrc1*^{3A} cells showed any TAR. However, upon stress, TAR was strongly induced only in *mrc1*^{3A} cells when transcription and replication progressed towards each other (IN). Of note, cells deficient in the *HOT1* transcription factor, which drives the induction of *STL1*, did not show TAR in *mrc1*^{3A} cells (Supplementary Fig. 13c). Therefore, phosphorylation of Mrc1 is a key event to prevent TAR. Correspondingly, osmostressed *mrc1*^{3A} cells showed an increase in Rad52 foci when compared to wild type (Fig. 4c), as expected for an increase in Rad52-engaged recombination events. This demonstrates that phosphorylation of Mrc1 by Hog1 is important for reducing recombination events at genomic loci.

The genomic instability in *mrc1*^{3A} cells upon osmostress does not render cells osmosensitive. Therefore, there is probably a mechanism that prevents the lethal accumulation of the molecular events responsible for genomic instability. Deletion of *RAD53* did not affect cell growth in the presence of osmostress; however, its deletion in combination with *mrc1* or *mrc1*^{3A} resulted in osmosensitive cells (Fig. 4d). Therefore, transcription–replication collisions occurring in the absence of a cell-cycle delay upon osmostress depends on the DNA-damage checkpoint to maintain viability.

Stress induces a rapid and transient response in gene expression[1,2] to maximize cell survival. However, when stress occurs during S phase, an increase of gene expression poses the risk of collision between the replication and transcription machineries, which can only be prevented if DNA replication is delayed while transcription lasts. To coordinate transcription and replication, a single signal transduction kinase, Hog1, is able to target both processes by inducing adaptive transcription while delaying replication (Supplementary Fig. 1).

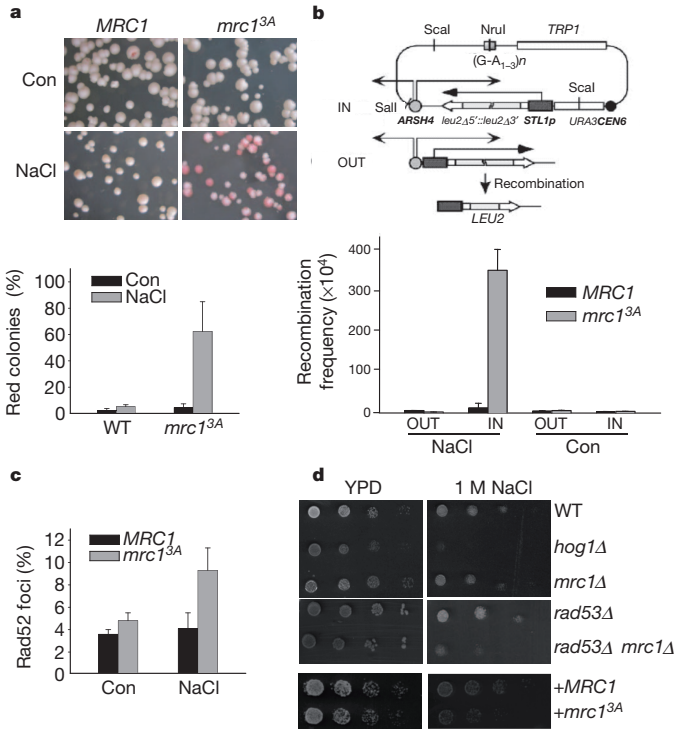


Figure 4 | Hog1–Mrc1 checkpoint prevents genomic instability upon osmostress. **a**, *mrc1*^{3A} cells show an increased frequency of plasmid loss upon osmostress. S-phase cells (wild type (WT) and *mrc1*^{3A}) were plated on YPD (Con) or YPD plus NaCl (osmostress), and red colonies were quantified. **b**, Schematic of the IN/OUT vectors used in the TAR assays. The graph shows TAR frequency in wild-type (*MRC1*) or *mrc1*^{3A} cells in absence (control) or presence of osmostress (NaCl). **c**, Rad52 foci in response to NaCl. Wild-type and *mrc1*^{3A} cells carrying *RAD52-YFP* were scored for foci (Con) or after 4 h of osmostress (NaCl). **a**–**c**, Data represent the mean and s.d. of four independent experiments. **d**, *mrc1*^{3A} cells are osmosensitive when the Rad53-dependent checkpoint pathway is impaired.

The known S-phase checkpoint is a surveillance mechanism that inhibits and protects chromosome replication in response to internal signals that arise from the replication process and DNA damage. We show that Hog1 and Mrc1 define an alternative S-phase checkpoint to integrate extracellular stimuli into DNA replication, ensuring proper coordination of transcription and replication upon stress. This coordination prevents the genomic instability that would otherwise be caused by collision of the two machineries. Remarkably, the two pathways share Mrc1 to delay S-phase progression. Mrc1 has two sets of phosphorylation sites, some targeted by Mec1 that lead to Rad53 activation and delay late origin firing, and some targeted by Hog1 that alter Cdc45 association to early origins and delay both early and late origin firing (Supplementary Fig. 15). Therefore, not only is the input signal received by Mrc1 different for Mec1 and Hog1, but the outcome of phosphorylation by the two kinases is also different, thus defining completely independent pathways from top to bottom.

There are many extracellular stimuli that require the rapid induction of a vast number of genes to maximize cell survival. These changes in gene expression need to be compatible with replication. This study provides insights about how cells can deal with high transcription rates while replicating, which might reflect a common need in eukaryotic cells.

METHODS SUMMARY

Genetic methods, strains and plasmids and biochemical experiments are described in Methods.

Full Methods and any associated references are available in the online version of the paper.

Received 4 June; accepted 15 October 2012.
Published online 25 November 2012.

1. López-Maury, L., Marguerat, S. & Bahler, J. Tuning gene expression to changing environments: from rapid responses to evolutionary adaptation. *Nature Rev. Genet.* **9,** 583–593 (2008).
2. de Nadal, E., Ammerer, G. & Posas, F. Controlling gene expression in response to stress. *Nature Rev. Genet.* **12,** 833–845 (2011).
3. Prado, F. & Aguilera, A. Impairment of replication fork progression mediates RNA polII transcription-associated recombination. *EMBO J.* **24,** 1267–1276 (2005).
4. Pomerantz, R. T. & O'Donnell, M. Direct restart of a replication fork stalled by a head-on RNA polymerase. *Science* **327,** 590–592 (2010).
5. Muers, M. Mutation: the perils of transcription. *Nature Rev. Genet.* **12,** 156 (2011).
6. de Nadal, E. & Posas, F. Multilayered control of gene expression by stress-activated protein kinases. *EMBO J.* **29,** 4–13 (2010).
7. Yaakov, G. *et al.* The stress-activated protein kinase Hog1 mediates S phase delay in response to osmostress. *Mol. Biol. Cell* **20,** 3572–3582 (2009).
8. Alcasabas, A. A. *et al.* Mrc1 transduces signals of DNA replication stress to activate Rad53. *Nature Cell Biol.* **3,** 958–965 (2001).
9. Osborn, A. J. & Elledge, S. J. Mrc1 is a replication fork component whose phosphorylation in response to DNA replication stress activates Rad53. *Genes Dev.* **17,** 1755–1767 (2003).
10. Katou, Y. *et al.* S-phase checkpoint proteins Tof1 and Mrc1 form a stable replication-pausing complex. *Nature* **424,** 1078–1083 (2003).
11. Tourrière, H. *et al.* Mrc1 and Tof1 promote replication fork progression and recovery independently of Rad53. *Mol. Cell* **19,** 699–706 (2005).
12. Weake, V. M. & Workman, J. L. Inducible gene expression: diverse regulatory mechanisms. *Nature Rev. Genet.* **11,** 426–437 (2010).
13. Chen, R. E. & Thorner, J. Function and regulation in MAPK signaling pathways: lessons learned from the yeast *Saccharomyces cerevisiae*. *Biochim. Biophys. Acta* **1773,** 1311–1340 (2007).
14. Hohmann, S. Osmotic stress signaling and osmoadaptation in yeasts. *Microbiol. Mol. Biol. Rev.* **66,** 300–372 (2002).
15. Escoté, X. *et al.* Hog1 mediates cell-cycle arrest in G1 phase by the dual targeting of Sic1. *Nature Cell Biol.* **6,** 997–1002 (2004).
16. Clotet, J. *et al.* Phosphorylation of Hsl1 by Hog1 leads to a G2 arrest essential for cell survival at high osmolarity. *EMBO J.* **25,** 2338–2346 (2006).
17. Clotet, J. & Posas, F. Control of cell cycle in response to osmostress: lessons from yeast. *Methods Enzymol.* **428,** 63–76 (2007).
18. Wurgler-Murphy, S. M. *et al.* Regulation of the *Saccharomyces cerevisiae* HOG1 mitogen-activated protein kinase by the PTP2 and PTP3 protein tyrosine phosphatases. *Mol. Cell. Biol.* **17,** 1289–1297 (1997).
19. Shirayama, M. *et al.* APC(Cdc20) promotes exit from mitosis by destroying the anaphase inhibitor Pds1 and cyclin Clb5. *Nature* **402,** 203–207 (1999).
20. Jacobson, M. D. *et al.* Testing cyclin specificity in the exit from mitosis. *Mol. Cell. Biol.* **20,** 4483–4493 (2000).
21. Aparicio, O. M., Weinstein, D. M. & Bell, S. P. Components and dynamics of DNA replication complexes in *S. cerevisiae*: redistribution of MCM proteins and Cdc45p during S phase. *Cell* **91,** 59–69 (1997).
22. Lou, H. *et al.* Mrc1 and DNA polymerase ε function together in linking DNA replication and the S phase checkpoint. *Mol. Cell* **32,** 106–117 (2008).
23. Aguilera, A. The connection between transcription and genomic instability. *EMBO J.* **21,** 195–201 (2002).
24. Aguilera, A. mRNA processing and genomic instability. *Nature Struct. Mol. Biol.* **12,** 737–738 (2005).
25. García-Rubio, M. *et al.* Different physiological relevance of yeast THO/TREX subunits in gene expression and genome integrity. *Mol. Genet. Genomics* **279,** 123–132 (2008).
26. Gonzalez-Barrera, S., Garcia-Rubio, M. & Aguilera, A. Transcription and double-strand breaks induce similar mitotic recombination events in *Saccharomyces cerevisiae*. *Genetics* **162,** 603–614 (2002).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank L. Subirana, S. Ovejas and A. Fernandez for technical support. This work was supported by grants from the Spanish Government (BIO2009-07762 and BFU2012-33503 to F.P., BFU2011-26722 to E.d.N., BFU2010-16372 to A.A., and Consolider Ingenio 2010 programme CSD2007-0015 to F.P. and A.A.) and FP7 UNICELLSYS grant (no. 201142) and the Fundación Marcelino Botín to F.P. F.P. and E.d.N. are recipients of an ICREA Acadèmia (Generalitat de Catalunya).

Author Contributions A.D. conducted most of the experiments. I.F.-A., S.B. and M.G.-R. worked on the analysis of replication. G.Y. initiated the studies. A.D., E.d.N., A.A., G.Y. and F.P. did the experimental designs. A.D., A.A., E.d.N. and F.P. wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to F.P. (francesc.posas@upf.edu).

METHODS

DNA plasmids and yeast strains. Plasmid pAD88 (*MRC1*) was obtained by cloning the *MRC1* open reading frame (ORF) and its promoter (323 bp upstream from the ORF) in the BamHI site of the episomal vector pRS415. pAD66 (*mrc1^{3A}*) and pAD79 (*mrc1^{3D}*) were obtained by sequential site-directed mutagenesis of pAD88. In these mutant constructs, Thr 169, Ser 125 and Ser 229 are mutated to Ala or Asp, respectively. To obtain the integrative constructs pAD74 (*MRC1*) and pAD75 (*mrc1^{3A}*), and pAD103 (*mrc1^{3D}*), ORFs and their promoters were isolated from pAD88, pAD66 and pAD79, respectively, and cloned into the pRS406 integrative vector. pAD57 (*MRC1*) was obtained by cloning *MRC1* in the BamHI site of the pGEX-6P-1 vector in-frame of the *GST* N-terminal tag. pAD67 (*mrc1^{3A}*) was obtained by cloning the *mrc1^{3A}* allele isolated from pAD66 by BamHI digestion and cloned into the BamHI site of the pGEX-6P-1 vector in-frame of the *GST* N-terminal tag. *STL1*-IN and *STL1*-OUT vectors are based on the GAL-IN and GAL-OUT vectors, respectively[3], and were modified by swapping the *GAL1* promoter with the Hog1-dependent *STL1* promoter. *STL1* expression is controlled by the Hot1 transcription factor. To detect Rad52 foci pWJ1344 carrying *RAD52-YFP*[27] was used. For the two-hybrid assay, pAD99 was obtained by cloning the full-length *MRC1* ORF in the bait vector pBTM116. pACTII-Hog1 was also used in the two-hybrid assay[28].

The strains used in this work were derived from W303 (*MAT*a, *his3 leu2 trp1 ura3 ade2 can1*) or its mutant *cdc7-4* (*MAT*a, *his3 leu2 trp1 ura3 ade2 can1*, *cdc7-ts4*), except for the genomic instability assay, in which Yph277 (ref. 29) was used. The *cdc7-4* allele was characterized previously[30]. The following strains were used to follow cell cycle progression by FACS and western blot: YAD93 (*MAT*a *his3 leu2 trp1 ura3 ade2 can1 cdc7-ts4 mrc1::KanMX*, *Clb5-TAP*) transformed with pRS415, pAD88 (pRS415-*MRC1*), pAD66 (pRS415-*mrc1^{3A}*) or pAD79 (pRS415-*mrc1^{3D}*), YAD125 (*cdc7-ts4 mrc1Δ*, pAD74 (pRS406-*MRC1*)) and YAD126 (*cdc7-ts4 mrc1Δ*, pAD75 (pRS406- *mrc1^{3A}*)). In the ChIP experiments we used the following strains: YAD127 (*cdc7-ts4 mrc1Δ*, pAD74 (pRS406-*MRC1*) *DPB2-HA*), YAD128 (*cdc7-ts4 mrc1Δ*, pAD74 (pRS406- *mrc1^{3A}*) *DPB2-HA*), YAD192 (*cdc7-ts4 mrc1Δ*, pAD74 (pRS406- *mrc1^{3D}*) *DPB2-HA*), YAD113 (*cdc7-ts4 mrc1Δ*, pAD74 (pRS406-*MRC1*) *CDC45-HA*), YAD179 (*cdc7-ts4 mrc1Δ*, pAD74 (pRS406- *mrc1^{3A}*) *CDC45-HA*), YAD190 (*sml1Δ rad53Δ CDC45-HA*), YAD155 (*cdc7-ts4 MRC1-HA*), YAD184 (*cdc7-ts4 mrc1Δ*, pAD74 (pRS406-*MRC1*) *POL2-HA*), YAD185 (*cdc7-ts4 mrc1Δ*, pAD74 (pRS406- *mrc1^{3A}*) *POL2-HA*). For immunoprecipitation assays, the strains YAD134 (*HOG1-HA MRC1-TAP*), YAD103 (*MRC1-TAP*), YAD122 (*POL2-HA MRC1-MYC*), YAD137 (*POL2-HA*) and YAD168 (*POL2-HA mrc1^{3A}-MYC*) were used to assay the interaction between DNA Pol2 and Mrc1 or *Mrc1^{3A}*. YAD144 (*sml1Δ mec1Δ MRC1-TAP*) and YAD105 (*MRC1-TAP hog1*) were used to assay the interaction between Hog1 and Mrc1. All tags (HA, TAP and Myc) were integrated at the C terminus of the gene (*CLB5*, *DPB2*, *MRC1*, *HOG1, POL2*, *CDC45*). YAD125 and YAD126 were also used for two-dimensional gel electrophoresis of replication forks and recombination assays to detect TAR. YAD192 (*cdc7-ts4 mrc1Δ*, pAD75 (pRS406- *mrc1^{3A}*) *hot1Δ::KANMX*) was also used in the TAR assays. In the latter case, these strains were transformed with the *STL1*-IN and *STL1*-OUT plasmids. For Rad52 foci experiments, the W303-1BR5 (ref. 31) and WMR5 (*MATα ade2-1 can1-100 his3-11,15 leu2-3,112 trp1-1 ura3-1 mrc1Δ::KanMX, KanMXΔ::NAT* pAD75 (pRS406 *mrc1^{3A}*)*::URA3 RAD5+*) strains were used. Osmosensitivity assays were performed with YAD143 (*sml1Δ rad53Δ mrc1Δ*) carrying the pRS415, pAD66 or pAD88 plasmids. For DNA combing, strains were modified to allow incorporation of exogenous BrdU into genomic DNA. Seven copies of the herpes simplex *TK* gene under the control of the yeast GDP promoter were inserted at the *URA3* locus. The resulting strains used in the DNA-combing experiments were: WRBMb-6C (*MAT*a *ade2-1 can1-100 his3-11,15 leu2-3,112 trp1-1 bar1Δ:: MRC1::TAP::KAN RRM3::FLAG::KAN ura3:: URA3/GDP-TK(7X)*) and WRB3Ab-6C (*MAT*a *ade2-1 can1-100 his3-11,15 leu2-3,112 trp1-1 bar1Δ:: mrc1^{3A}::TAP::KAN RRM3::FLAG::KAN ura3:: URA3/GDP-TK(7X)*).

Expression and purification of recombinant proteins. *Escherichia coli* bacteria were grown at 37 °C to an $OD_{600\,nm}$ of 0.5. Then GST-tagged proteins were induced for 6 h by adding 1 mM isopropylthiogalactoside (IPTG) at 25 °C. After induction cells were collected by centrifugation and resuspended in 1/50 volume of STET 1× buffer (100 mM NaCl, 10 mM Tris-HCl pH 8.0, 10 mM EDTA pH 8.0, 5% Triton X-100 supplemented with 2 mM dithiothreitol (DTT) and 1 mM phenylmethylsulphonyl fluoride (PMSF), 1 mM benzamidine, 200 μg ml^{-1} leupeptin and 200 μg ml^{-1} pepstatin). Cells were lysed in ice-cold sonication and cleared by high-speed centrifugation. GST-fused proteins were pulled down from supernatants with 300 μl of 4B gluthatione-sepharose beads (GE Healthcare, 50% slurry equilibrated with STET) by mixing for 90 min at 4 °C. The gluthatione-sepharose beads were collected by brief centrifugation and washed four times in STET 1× buffer and two times in 50 mM Tris-HCl pH 8.0 buffer supplemented with 2 mM DTT. The GST-fused proteins were then eluted in 200 μl of 50 mM Tris-HCl pH 8.0 buffer supplemented with 2 mM DTT and 10 mM reduced gluthatione (Sigma) by rotating for 20 min at 4 °C and stored at −80 °C.

Kinase assays. *GST-MRC1* and *GST-mrc1^{3A}* were expressed in *E. coli* and purified. Eluted proteins were mixed with 1 μg of purified GST–Hog1 (pre-activated with *GST-PBS2^{EE}*) in 1× kinase buffer (50 mM Tris-HCl pH 7.5, 10 mM $MgCl_2$, 2 mM DTT) supplemented with 100 μM cold ATP and γ^{32}-ATP (5 μM). The mix was incubated for 20 min at 37 °C and samples were resolved using SDS–PAGE, the proteins were visualized by Coomassie staining, and phosphorylation was detected using X-OMAT (Kodak) films.

Western blotting. TCA protein extracts were resolved using SDS–PAGE. Total amounts of the indicated proteins were detected by immunobloting using the indicated antibodies with the ECL detection reagent (Amersham). For quantifications, the corresponding films were scanned at 16 bits per channel and quantified using QuantityOne Application Software version 4.6.1.

Immunoprecipitation assays. The selected strains were grown to mid-log exponential phase ($OD_{660\,nm}$ = 1) and stressed or not with 0.4 M NaCl for 7 min at 25 °C. Cells were collected (400 ml per condition) and kept at −80 °C, and pellets were resuspended in 2 ml of lysis buffer (45 mM HEPES-KOH pH 7.2, 150 mM NaCl, 1 mM EDTA, 10% glycerol, 0.2% NP40) containing the whole cocktail of antiproteases and phosphatase inhibitors. An equal volume of glass beads (of diameter 0.5 mm) was added, and cells were broken by vortexing at 4 °C. The whole extract was clarified by centrifuging for 10 min at 10,500*g* and an aliquot was taken for further analysis. 3–7 mg of the extracts were incubated overnight with anti-Myc monoclonal antibody (9E10) or with anti-HA affinity matrix (Roche), and the beads were washed five times with the lysis buffer. Antibody-bound fractions and the corresponding whole-cell extract were boiled in SDS-containing sample buffer and loaded in 8% acrylamide gels.

Growth conditions and FACS analyses. Overnight cultures were diluted to an $OD_{660\,nm}$ of 0.3 and grown for 3 h at 25 °C in YPD. *cdc7-4* cells were synchronized at the onset of S phase in two step. First, they were incubated with α-factor for 2 h at 25 °C (40 μg ml^{-1}), washed and then released at 37 °C (incubated for 2 h) to block the cell cycle at the beginning of S phase, before DNA replication starts. Second, they were released from the temperature block, and the cells were allowed to progress into S phase at 25 °C, in YPD, supplemented or not with 0.4 M NaCl or 0.8 M sorbitol. For the *GAL1::Pbs2^{DD}* overexpression experiments, cells were grown overnight in synthetic, minimal medium (SD) with 2% raffinose. The *GAL1* promoter was induced with 2% galactose after 90 min at 37 °C, and left for an extra 90 min before release from the temperature block. For FACS, cells were fixed in ethanol, treated overnight with RNase A at 37 °C in 50 mM sodium citrate, stained with propidium iodide and analysed in a FACScan flow cytometer (Becton Dickinson). A total of 10,000 cells were analysed and the population of G_1 quantified for each time point using WinMDI 2.9.

ChIP assays. In ChIP experiments, cells were treated as explained above, but they were released into S phase at 16 °C to assess Dpb2–HA and Mrc1–HA binding onto DNA sequences. Cells (50 ml per point) from cell-cycle time courses were collected at $OD_{660\,nm}$ = 0.7 were treated with 1% formaldehyde for 20 min at room temperature. Glycine was added to a final concentration of 330 mM and the incubation continued for 15 min. Cells were collected, washed four times with cold TBS (20 mM Tris-HCl, pH 7.5, 150 mM NaCl), and kept at −20 °C for further processing. Cell pellets were resuspended in 0.3 ml cold lysis buffer (50 mM HEPES-KOH, pH 7.5, 150 mM NaCl, 1 mM EDTA, 0.1% sodium deoxycholate, 0.1% SDS, 1 mM PMSF) supplemented with 1% Triton-X 100. An equal volume of glass beads (of diameter 0.5 mm) was added, and cells were disrupted by vortexing (with vortex gene) for 13 min on ice. The lysate was diluted into 0.6 ml lysis buffer, and the glass beads were discarded. The crosslinked chromatin was sonicated to yield an average DNA fragment size of 350 bp (range, 100–850 bp). Finally, the sample was clarified by centrifugation at 15,500*g* for 5 min at 4 °C. Chromatin solution (600 μl) was incubated with 50 μl anti-HA monoclonal antibody pre-coupled to anti-mouse IgG-conjugated paramagnetic beads (Dynabeads M-450, Dynal). After 90 min at 4 °C on a rotator, beads were washed twice for 4 min in 1 ml lysis buffer, twice in 1 ml lysis buffer with 500 mM NaCl, twice in 1 ml washing buffer (10 mM Tris-HCl pH 8.0, 0.25 M LiCl, 1 mM EDTA, 0.5% N-P40, 0.5% sodium deoxycholate) and once in 1 ml TE (10 mM Tris-HCl pH 8.0, 1 mM EDTA). Immunoprecipitated material was eluted twice from the beads by heating for 10 min at 65 °C in 50 μl elution buffer (25 mM Tris-HCl pH 7.5, 10 mM EDTA, 0.5% SDS). To reverse crosslinks, samples were adjusted to 0.3 ml with elution buffer and incubated overnight at 65 °C. After extraction with phenol-chloroform-isoamyl alcohol and chloroform, DNA was ethanol precipitated for 4 h at −20 °C in the presence of 20 μg glycogen, and resuspended in 30 μl of TE buffer. For real-time PCR, oligonucleotides for the *ARS305* and *ARS501* origin sequences, and for sequences located at 8 kb or 17 kb downstream of the *ARS305* origin were used. As an internal PCR control, oligonucleotides for telomeric regions were used.

Two-dimensional gel electrophoresis. Cells growing exponentially in YPD at 25 °C were synchronized in G_1 phase with α-factor (3 μg ml^{-1}) for 2 h 30 min, washed free of α-factor and immediately shifted to 37 °C for 2 h to synchronize cells in S phase. HU was added at a final concentration of 40 mM 30 min before the shift to 25 °C. The culture was treated with 0.4 M NaCl and shifted back to 25 °C. Total DNA from 50 ml of mid-log-phase cells was isolated as previously described[31]. DNA was restricted with EcoRV and/or HindIII. Two-dimensional gel electrophoresis was performed as previously described[31]. For hybridization, the coordinates of the α^{32}P PCR probes used were 32170-35159 to detect the HindIII-HindIII fragment and 37883-41883 for the EcoRV-HindIII fragment of *ARS305* on chromosome III. Signals were quantified using a Phosphorimager Fujifilm FLA-5100 and the ImageGauge program. For two-dimensional gel quantification signals were quantified using a Phosphorimager Fujifilm FLA-5100 and the Image Gauge program. The ratio of bubbles (percentage) was normalized with respect to the total amount of replicating molecules to determine the relative delay of replication fork initiation. Quantification of the replicative intermediates in the fragment adjacent to *ARS305* was normalized with respect to the total amount of signals present in the gel, including linear monomers (*n*).

Rad52 foci. Spontaneous and NaCl-induced Rad52–YFP foci were counted in 4′,6-diamidino-2-phenylindole (DAPI)-stained nuclei from S–G_2 mid-log cells bearing the plasmid pWJ1344. For NaCl treatment, cells were incubated in medium supplemented with NaCl at a final concentration of 0.8 M for an additional 4 h before the quantification of the Rad52–YFP foci, which were visualized with a Leica DC 350F microscope.

Genomic instability assay. Overnight cultures of the Yph277 strain transformed with plasmids harbouring the wild type or the *mrc1*^{3A} mutant were diluted to an $OD_{660\,nm}$ of 0.3, grown in YPD until an $OD_{660\,nm}$ of 0.5, and synchronized with α-factor for 90 min at 30 °C (40 μg ml^{-1}). Cells were washed and released into S phase in YPD for 30 min and 200 μl of cells (1/1,000 diluted) of each culture were plated in YPD or YPD with 0.8 M NaCl, and incubated at 30 °C for 3 days and left at 4 °C.

Recombination assays. *cdc7-4* cells (wild type and *mrc1*^{3A}) harbouring *STL1*-IN or *STL1*-OUT vectors (which contain the origin of replication ARS209 or ARSH4 as named in the scheme) were synchronized as described above and 200 μl of S-phase cells (1/100 diluted) were plate in SD Trp$^-$ plates (control plates) or in SD Trp$^-$ Leu$^-$ plates (recombination plates), supplemented or not with 0.8 M NaCl.

Two-hybrid assay and β-galactosidase assay. The two-hybrid analysis was carried out by using pACTII as the activation domain plasmid and pBTM116-LexA as the DNA-binding domain plasmid. The plasmid LexA-Mrc1 was cotransformed with pACT-Hog1 using the L40 (*MATa trp1 leu2 his3 LYS::lexA-HIS3 URA3::lexA-LacZ*) reporter strain. We performed the corresponding negative controls (using the empty vectors) and positive controls (the well-known interactors LexA-Raf and pACT-Ras) in parallel. Positive clones were selected and further tested for β-galactosidase activity as follows. The transformed yeast strains were grown selectively until mid-log phase in the appropriate selective liquid media and then diluted in YPD for 3 h. Logarithmically growing cells ($OD_{660\,nm}$ = 0.5–0.8) were treated with 0.4 M of NaCl for 35 min and permeabilized by ethanol-toluene treatment, and β-galactosidase activity was determined as light intensity at $OD_{415\,nm}$ after the addition of *o*-nitrophenil-β-D-galactoside (ONPG) substrate.

DNA combing. Mid-log phase cells were synchronized with α-factor and incubated for 20 min in 40 mM HU before adding 200 μg ml^{-1} BrdU (Sigma) and 0.4 M NaCl. DNA combing was performed as described[32]. DNA fibres were extracted in agarose plugs after 30 and 45 min of BrdU labelling, corresponding to 50 and 65 min after one release, and were stretched on silanized coverslips. DNA molecules were counterstained with an anti-ssDNA antibody (MAB3868, Chemicon; 1/100) and a goat anti-mouse antibody coupled to Alexa 546 (A11030, Molecular Probes, 1/50). BrdU was detected with BU1/75 (AbCys, 1/20) anti-BrdU antibody. DNA fibres were analysed in a Leica DM6000 microscope equipped with a DFC390 camera (Leica). Data acquisition was performed with LAS AF (Leica). Velocity was estimated as the difference between the median values of the track length of the 45 and 30 min time points divided by 15. In asynchronous culture the combing assay was performed by incubating mid-log phase cells for 45 min in medium with 200 μg ml^{-1} BrdU with or without 0.4 M NaCl.

27. Lisby, M., Mortensen, U. H. & Rothstein, R. Colocalization of multiple DNA double-strand breaks at a single Rad52 repair centre. *Nature Cell Biol.* **5,** 572–577 (2003).
28. Nadal, E., Casadome, L. & Posas, F. Targeting the MEF2-like transcription factor Smp1 by the stress-activated Hog1 mitogen-activated protein kinase. *Mol. Cell. Biol.* **23,** 229–237 (2003).
29. Spencer, F. *et al.* Mitotic chromosome transmission fidelity mutants in *Saccharomyces cerevisiae*. *Genetics* **124,** 237–249 (1990).
30. Bousset, K. & Diffley, J. F. The Cdc7 protein kinase is required for origin firing during S phase. *Genes Dev.* **12,** 480–490 (1998).
31. Moriel-Carretero, M. & Aguilera, A. A postincision-deficient TFIIH causes replication fork breakage and uncovers alternative Rad51- or Pol32-mediated restart mechanisms. *Mol. Cell* **37,** 690–701 (2010).
32. Bianco, J. N. *et al.* Analysis of DNA replication profiles in budding yeast and mammalian cells using DNA combing. *Methods* **57,** 149–157 (2012).

LETTER

doi:10.1038/nature11658

DNA-repair scaffolds dampen checkpoint signalling by counteracting the adaptor Rad9

Patrice Y. Ohouo[1], Francisco M. Bastos de Oliveira[1], Yi Liu[1], Chu Jian Ma[1] & Marcus B. Smolka[1]

In response to genotoxic stress, a transient arrest in cell-cycle progression enforced by the DNA-damage checkpoint (DDC) signalling pathway positively contributes to genome maintenance[1]. Because hyperactivated DDC signalling can lead to a persistent and detrimental cell-cycle arrest[2,3], cells must tightly regulate the activity of the kinases involved in this pathway. Despite their importance, the mechanisms for monitoring and modulating DDC signalling are not fully understood. Here we show that the DNA-repair scaffolding proteins Slx4 and Rtt107 prevent the aberrant hyperactivation of DDC signalling by lesions that are generated during DNA replication in *Saccharomyces cerevisiae*. On replication stress, cells lacking Slx4 or Rtt107 show hyperactivation of the downstream DDC kinase Rad53, whereas activation of the upstream DDC kinase Mec1 remains normal. An Slx4–Rtt107 complex counteracts the checkpoint adaptor Rad9 by physically interacting with Dpb11 and phosphorylated histone H2A, two positive regulators of Rad9-dependent Rad53 activation. A decrease in DDC signalling results from hypomorphic mutations in *RAD53* and *H2A* and rescues the hypersensitivity to replication stress of cells lacking Slx4 or Rtt107. We propose that the Slx4–Rtt107 complex modulates Rad53 activation by a competition-based mechanism that balances the engagement of Rad9 at replication-induced lesions. Our findings show that DDC signalling is monitored and modulated through the direct action of DNA-repair factors.

Slx4 is an evolutionarily conserved DNA-repair scaffolding protein that is important for the cellular response to exogenous DNA-damaging agents[4–7], and mutations in human *SLX4* were recently linked to Fanconi anaemia[8,9]. In *S. cerevisiae* (budding yeast), cells that lack *SLX4* (*slx4*Δ cells) are highly sensitive to methyl methanesulphonate (MMS)[7], a DNA-alkylating agent that blocks replication and induces the DDC pathway. While investigating the activation status of the *S. cerevisiae* DDC kinase Rad53 in *slx4*Δ cells, we noted that MMS treatment leads to hyperphosphorylation, and thus hyperactivation, of Rad53 compared with in wild-type cells (Fig. 1a and Supplementary Fig. 1a), which is consistent with a previous report[10]. However, the phosphorylation of histone H2A (also known as Hta1 and Hta2), a substrate of the upstream DDC kinase Mec1, at serine 129 (here referred to as $H2A^{PS129}$) was not increased in *slx4*Δ cells (Fig. 1a, lower panel). These results suggest that the hyperactivation of Rad53 in *slx4*Δ cells is not caused by increased damage-induced Mec1 signalling but by improper downstream regulation of Rad53 activation. To test this possibility, we compared the phosphoproteome of wild-type and *slx4*Δ cells after MMS treatment, using quantitative mass spectrometry. Although most of the detected Mec1 targets were phosphorylated to the same extent in both cell types, Rad53-dependent phosphorylation was significantly higher in *slx4*Δ cells than in wild-type cells (Supplementary Fig. 1b), further supporting the idea that Slx4 has a role in specifically blocking Rad53 hyperactivation. Because the activation of Rad53 in response to MMS mostly depends on the checkpoint adaptor Rad9 (Fig. 1b and Supplementary Fig. 2), Slx4 probably counteracts Rad9-dependent Rad53 activation.

To test whether the sensitivity of *slx4*Δ cells to MMS is caused mostly by aberrant Rad53 hyperactivation, we used hypomorphic alleles of *rad53* that result in lower Rad53 activation, reasoning that these alleles would rescue the MMS sensitivity of *slx4*Δ cells. Rad53 has two FHA domains, which bind to phosphorylated Rad9 in a redundant manner and mediate Rad53 activation[11] (Fig. 1c). Mutations in the FHA2 domain promote a stronger reduction in MMS-induced Rad53 activation than do mutations in the FHA1 domain[12]. Whereas a mutation (R70A, in which arginine is substituted with alanine at amino acid 70) in the FHA1 domain of Rad53 had no effect on the MMS sensitivity of *slx4*Δ cells, a mutation (R605A) in the FHA2 domain reduced the MMS sensitivity of *slx4*Δ cells (Fig. 1d). Consistent with our hypothesis that Rad53 hyperactivation is the cause of the MMS sensitivity of *slx4*Δ cells, mutation of the FHA2 domain resulted in a decrease in Rad53 activation in *slx4*Δ cells to a level similar to that of wild-type cells (Fig. 1e). Collectively, these results suggest that Slx4 has a crucial role in preventing excessive Rad9-dependent activation of Rad53 (Fig. 1f). The levels of MMS used here require that cells pass through the S phase of the cell cycle for Rad53 to become active[13]; therefore, our results suggest that Slx4 counteracts the Rad9–Rad53 pathway in response to replication-induced lesions. The finding that combined deletion of the *SLX1* and *RAD1* genes, which encode nucleases that are known to associate with Slx4, leads to lower MMS sensitivity and Rad53 activation than in *slx4*Δ cells (Supplementary Fig. 4) supports a nuclease-independent function for Slx4 during the cellular response to MMS-induced replication stress.

We recently reported that on replication stress, Slx4 binds to Dpb11 (ref. 14), a replication factor that is involved in DDC activation[15,16]. Because Dpb11 binds to Rad9 and positively regulates Rad9-dependent Rad53 activation[17,18], we assessed whether the Slx4–Dpb11 interaction has a role in counteracting Rad53 activation during MMS treatment. We previously showed that phosphorylation of Slx4 by Mec1 mediates the interaction of Slx4 with Dpb11 and that an Slx4 mutant lacking seven Mec1 consensus phosphorylation sites (Slx4-7MUT) cannot stably interact with Dpb11 (ref. 14). Rad53 is hyperactivated in *slx4-7MUT* cells (Fig. 2a), supporting a model in which the Slx4–Dpb11 interaction is important for preventing Rad53 hyperactivation.

Next, we tested whether the Slx4–Dpb11 interaction inhibits the ability of Rad9 to bind to Dpb11 in wild-type and *slx4*Δ cells. Deletion of *SLX4* leads to a significant increase in the MMS-induced interaction between Dpb11 and Rad9 (Fig. 2b and Supplementary Fig. 5), suggesting that Slx4 and Rad9 compete for Dpb11 binding. Dpb11 contains two pairs of BRCT domains, which bind to phosphorylated motifs. We found that recombinant BRCT domains 1 and 2 ($BRCT^{1/2}$) of Dpb11 can bind phosphorylated Slx4 and phosphorylated Rad9 from MMS-treated *S. cerevisiae* lysates (Fig. 2c). This finding is consistent with a model in which Slx4 and Rad9 compete for $BRCT^{1/2}$ binding.

Dpb11 $BRCT^{1/2}$ was previously shown to interact directly with cyclin-dependent kinase (CDK)-dependent phosphorylation sites in Sld3, thereby initiating DNA replication[19,20]. To better understand the mechanism of the Slx4–Dpb11 interaction, we searched for a

[1]Department of Molecular Biology and Genetics, Weill Institute for Cell and Molecular Biology, Cornell University, Ithaca, New York 14853, USA.

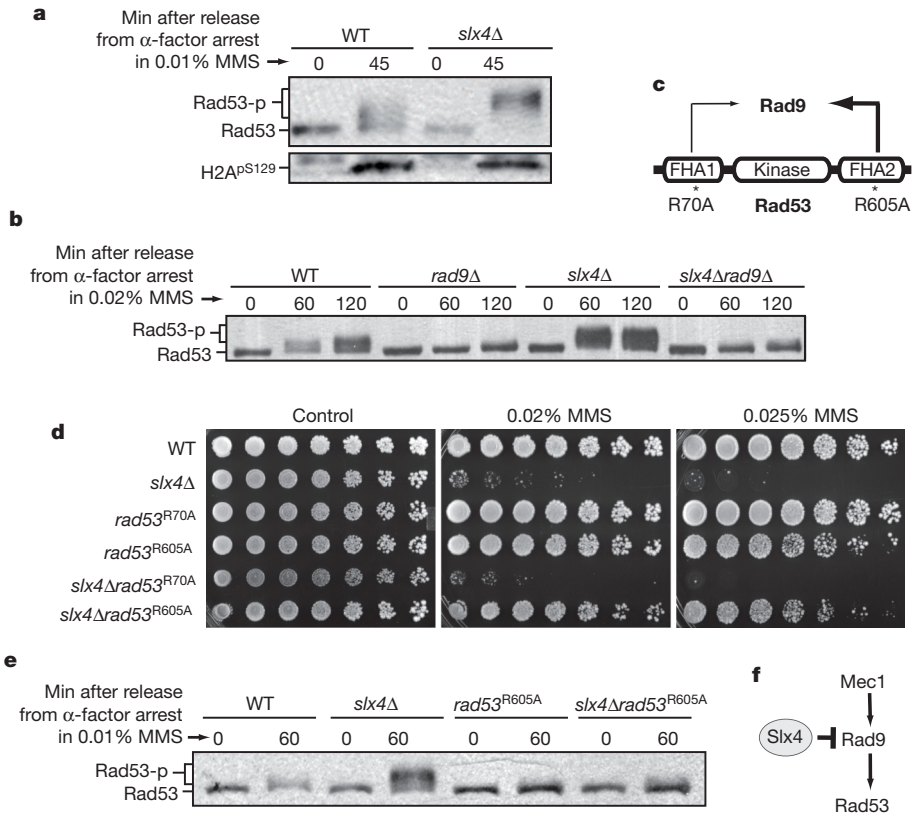


Figure 1 | Slx4 counteracts Rad9-dependent Rad53 activation. **a**, Western blot showing phosphorylation of haemagglutinin (HA)-tagged Rad53 and histone $H2A^{pS129}$ after MMS treatment. **b**, Western blot showing the phosphorylation status of Rad53–HA in the indicated strains after MMS treatment. **c**, Schematic representation of the Rad53 protein and its interaction with Rad9. **d**, MMS sensitivity assay of strains containing the indicated Flag-tagged mutant Rad53 proteins. Similar results were obtained with strains containing untagged Rad53 (Supplementary Fig. 3). **e**, Western blot showing the phosphorylation status of Rad53–Flag in the indicated strains after MMS treatment. **f**, Model for the role of Slx4 in uncoupling Rad53 activation from Mec1 signalling, through counteracting Rad9. Rad53-p, phosphorylated Rad53; WT, wild type.

CDK-targeting motif in Slx4 that resembles the Dpb11-binding region in Sld3. Notably, Slx4 contains proline-directed phosphorylation sites that align well with serine 600 (S600) and S622 of Sld3, which are important for the Dpb11–Sld3 interaction[19,20] (Supplementary Fig. 6). We tested the importance of the proline-directed sites within the Sld3-like region of Slx4 and found that S486 of Slx4 is crucial for the MMS-induced interaction with Dpb11 (Fig. 2d). In addition, three canonical Mec1 phosphorylation sites (S/T-Q sites) that are important for mediating a robust Slx4–Dpb11 interaction[14] are located in or near the Sld3-like Dpb11-binding region in Slx4. We conclude that the Slx4–Dpb11 interaction is mediated by the coordinated action of both Mec1 and a proline-directed kinase at a motif that is probably targeted by BRCT[1/2] of Dpb11. Although we detected residual binding between Dpb11 and Slx4-7MUT, we did not detect an interaction between Dpb11 and $Slx4^{S486A}$, suggesting that the phosphorylation of S486 has a more important role in mediating the Slx4–Dpb11 interaction than the Mec1 consensus phosphorylation sites. This is also supported by the finding that $slx4^{S486A}$ leads to Rad53 hyperactivation (Fig. 2e) and to a higher MMS sensitivity than does *slx4-7MUT* (Fig. 2f). Interestingly, previous reports showed that proline-directed sites in Rad9 are also important for the Dpb11–Rad9 interaction[17,18] (Supplementary Fig. 7).

To further test the model that Slx4 antagonizes Rad53 activation by binding to Dpb11 and outcompeting Rad9, we overexpressed Slx4 using an *ADH1* or a *TDH3* promoter and monitored different steps of checkpoint activation. In an early step in checkpoint activation, Rad9 assembles into a ternary complex with Dpb11 and Mec1 and is hyperphosphorylated by Mec1 (ref. 18). We first monitored the phosphorylation status of Rad9 in cells expressing Slx4 from its endogenous promoter and in cells overexpressing Slx4. Overexpression of wild-type Slx4 but not S486A mutated Slx4 significantly inhibited the MMS-induced hyperphosphorylation of Rad9 (Fig. 2g), as shown by the strong decrease in the slower migrating band at 45 and 60 min after release from α-factor arrest. We then monitored the effect of Slx4 overexpression on Rad53 activation in response to MMS treatment (Supplementary Fig. 8a). Although overexpression of Slx4 leads to a small but consistent decrease in Rad53 activation early in the response, we did not observe the same effect at later time points. We speculated that after 40 min, the Rad53 activation that was observed in Slx4-overexpressing cells was being mediated by a parallel mechanism based on Dot1-mediated histone H3K79 methylation, which can promote the recruitment of Rad9 to lesion sites independently of Dpb11 (ref. 21). We therefore monitored Rad53 activation in cells lacking *DOT1* and found that Slx4 overexpression significantly reduces Rad53 activation also at later time points (Fig. 2h); this effect depends on S486 of Slx4 (Supplementary Fig. 8b, c). Next, we determined whether Slx4 overexpression specifically disrupts the Rad9–Dpb11 interaction. We detected hyperphosphorylated Rad9 in a Dpb11 pull-down from cells expressing Slx4 from its endogenous promoter, but we did not detect Rad9 in a Dpb11 pull-down using cells that overexpressed Slx4 (Fig. 2i). This effect was specific for S486 of Slx4. These findings support our model that Slx4 counteracts DDC signalling by binding to Dpb11 and preventing its stable interaction with Rad9 (Fig. 2j).

Slx4 forms a tight complex with the BRCT-domain-containing protein Rtt107, which is a DNA-repair scaffold that stabilizes the Slx4–Dpb11 interaction[14]. Analysis of the phosphorylation status of Rad53 in *rtt107*Δ cells showed that DDC signalling is hyperactivated (Fig. 3a), suggesting that Rtt107 has a similar function to Slx4 in counteracting

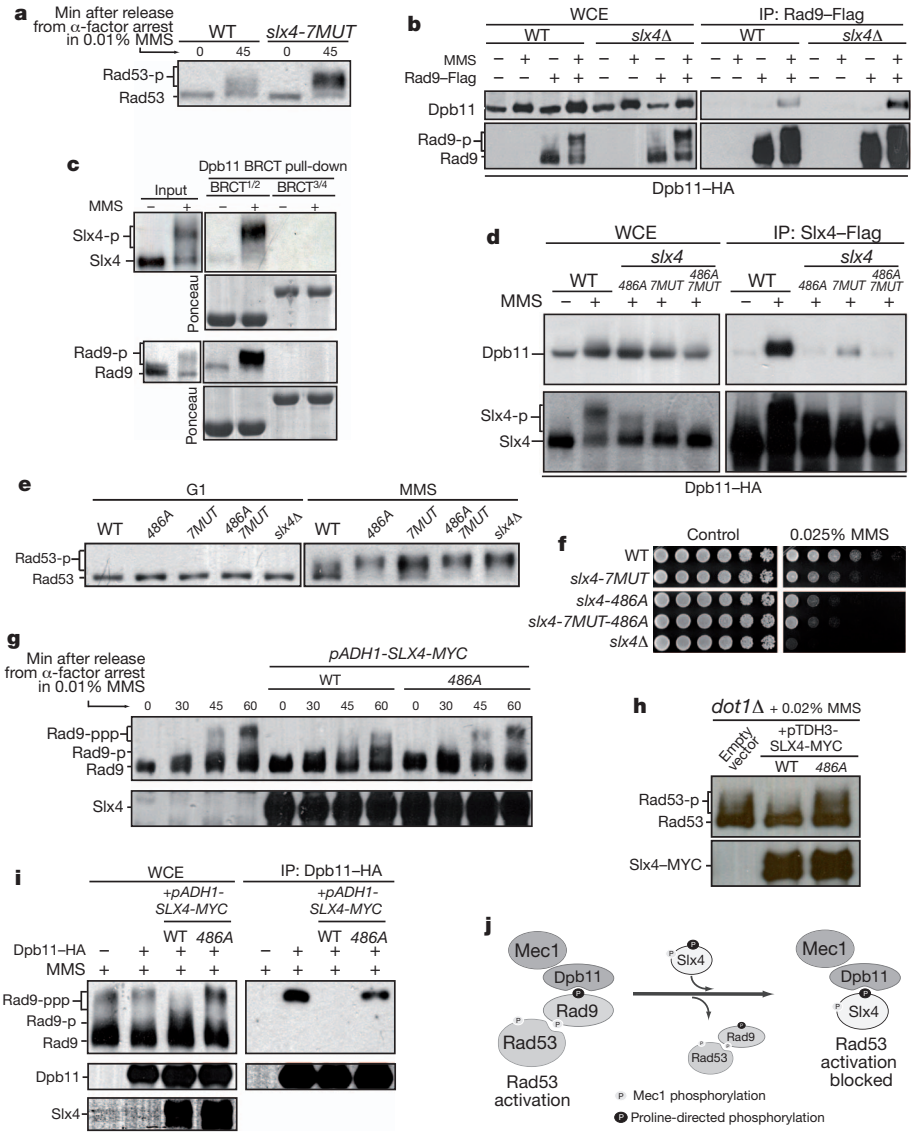


Figure 2 | Slx4 binding to Dpb11 counteracts the Dpb11–Rad9 interaction and Rad53 activation. **a**, Western blot showing the phosphorylation status of Rad53–HA in the indicated strains after MMS treatment. **b**, Co-immunoprecipitation (co-IP) of Dpb11 and Rad9 (see also Supplementary Fig. 5); input levels and phosphorylation status of Rad9 are shown. **c**, Pull-down of Slx4–Flag or Rad9–HA from *S. cerevisiae* lysates using recombinant $BRCT^{1/2}$ and $BRCT^{3/4}$ from Dpb11; ponceau staining of the membrane is also shown. **d**, Interaction of the indicated Flag-tagged Slx4 mutants with Dpb11–HA; input levels and phosphorylation status of Slx4 are shown. **e**, Rad53 phosphorylation status in the indicated *slx4* mutants. The experiment was performed as described in Fig. 1a. **f**, MMS sensitivity assay on the indicated *slx4* mutants. **g**, Rad9 phosphorylation status in cells expressing WT or mutant Slx4 from the endogenous *SLX4* promoter or the *ADH1* promoter. **h**, Rad53 phosphorylation status in *dot1*Δ cells expressing Slx4 from the endogenous *SLX4* promoter or the *TDH3* promoter. Cells were arrested in α-factor and released for 45 min in medium containing MMS. **i**, Co-IP of Dpb11 and Rad9. **j**, Model for the mechanism by which Slx4 counteracts Rad53 activation. G1, G1 cell-cycle phase; Rad9-p, hypophosphorylated Rad9; Rad9-ppp, hyperphosphorylated Rad9; WCE, whole cell extract.

Rad53 activation. Rtt107 recognizes $H2A^{pS129}$ through its carboxy-terminal pair of BRCT domains ($BRCT^{5/6}$), and this region is essential for the role of Rtt107 in the MMS response[22] (Supplementary Figs 9 and 10). Furthermore, the S129A mutation (*h2a-S129A*) abrogated Slx4 phosphorylation on MMS treatment (Fig. 3b), suggesting that the recruitment of Rtt107 to $H2A^{pS129}$ occurs upstream of Slx4–Dpb11 complex formation. Because $H2A^{pS129}$ can function as a positive regulator of Rad9-dependent Rad53 activation[23], we reasoned that the hyperactivation of Rad53 in *rtt107*Δ cells could be suppressed by *h2a-S129A*. *rtt107*Δ cells expressing the H2A(S129A) mutant had lower Rad53 activation and MMS sensitivity than did *rtt107*Δ cells expressing wild-type H2A (Fig. 3c, d). Disruption of the Rad9–H2A interaction by mutation of the C-terminal pair of BRCT domains of Rad9 (K1088M) partially rescued the MMS sensitivity of *rtt107*Δ cells (Supplementary Fig. 11a). A similar rescue was also observed when overexpressing $BRCT^{5/6}$ of Rtt107 (Supplementary Fig. 11b). Taken together, these results show that, like Slx4, Rtt107 also counteracts DDC signalling. The anti-checkpoint function of Rtt107 depends on its recognition of $H2A^{pS129}$, a step that is required subsequently for the assembly of the Slx4–Dpb11 complex. Thus, the Slx4–Rtt107 complex functions as a negative regulator of activation of the kinase Rad53 by physically interacting with two positive regulators of the adaptor Rad9: Dpb11 and $H2A^{pS129}$.

To further test the anti-checkpoint function of the Slx4–Rtt107 complex, we developed an alternative experimental set-up in which the presence of Slx4 would sensitize cells to replication stress by decreasing DDC signalling below the level required for a proper cellular response. For this set-up, we used cells lacking Mrc1, which is a checkpoint adaptor that works in parallel with Rad9. Mrc1 mediates Rad53 activation at stalled replication forks and therefore has a more

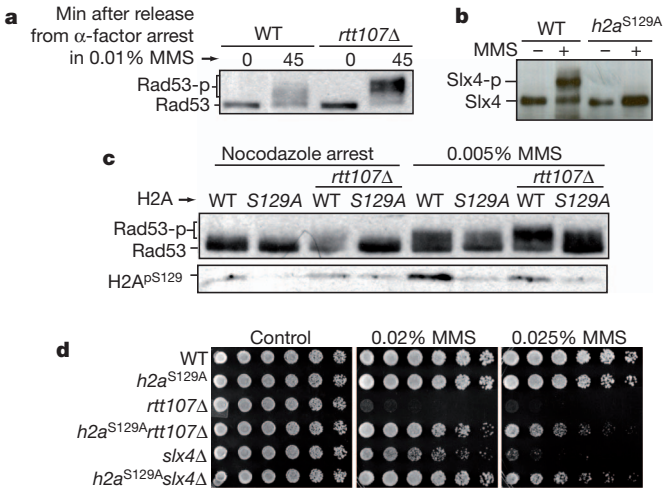


Figure 3 | Rtt107 counteracts Rad9-dependent Rad53 activation by binding to phosphorylated histone H2A. **a**, Western blot showing the phosphorylation status of Rad53–HA in the indicated strains after MMS treatment. **b**, Western blot showing Slx4–Flag phosphorylation status. **c**, Western blot showing phosphorylation status of Rad53 and H2A^{pS129}. Indicated strains expressing HA-tagged Rad53 were arrested in nocodazole and released for 60 min in medium containing MMS. **d**, MMS sensitivity assay on the indicated mutants.

central role than Rad9 in Rad53 activation in response to hydroxyurea[24]. On hydroxyurea treatment, *mrc1*Δ cells rely solely on Rad9 to activate Rad53. The interaction of Slx4 with Dpb11 was enhanced in *mrc1*Δ cells (Fig. 4a), especially in response to hydroxyurea, suggesting that the Slx4–Rtt107 complex actively counteracts Rad9 in *mrc1*Δ cells (Fig. 4b). Consistent with our finding that the Slx4–Dpb11 interaction requires H2A^{pS129}, high levels of H2A^{pS129} accumulated close to the origins of replication in *mrc1*Δ cells after hydroxyurea treatment (Fig. 4c). Because we expect that the kinase activity of Rad53 is limiting in *mrc1*Δ cells, we reasoned that deletion of *SLX4* would be beneficial in hydroxyurea-treated *mrc1*Δ cells, allowing more activation of Rad53 through Rad9. Consistent with this idea, *mrc1*Δ*slx4*Δ cells had a significantly higher hydroxyurea resistance than *mrc1*Δ cells (Fig. 4d), and this resistance correlated with higher Rad53 activation (Fig. 4e). Taken together, these results show that Slx4 sensitizes *mrc1*Δ cells to replication stress, and they provide strong additional evidence that the Slx4–Rtt107 complex counteracts Rad9 in response to replication-induced lesions (see detailed model in Supplementary Fig. 12). Because this function of the Slx4–Rtt107 complex depends on Mec1-dependent phosphorylation, we propose that Slx4–Rtt107 is involved in a mechanism that we have named DAMP — dampens checkpoint adaptor-mediated phospho-signalling — by which DDC signalling self-monitors its activation state. We speculate that by uncoupling upstream Mec1 signalling from downstream Rad53 activation, DAMP could allow Mec1 to maintain control over specific effectors, such as repair enzymes, without an aberrant arrest in cell-cycle progression.

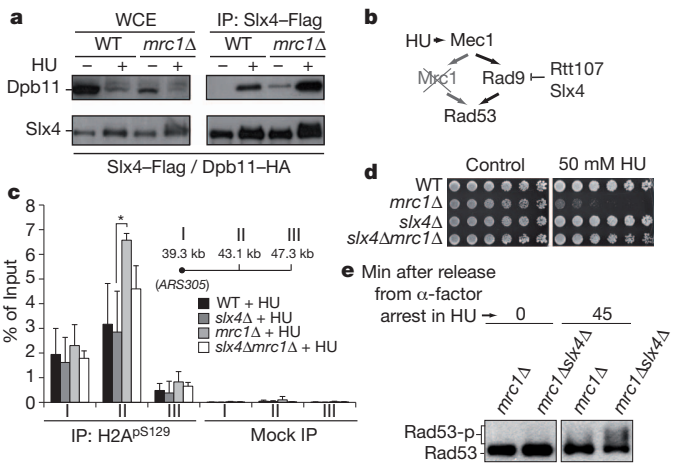


Figure 4 | Slx4 sensitizes *mrc1*Δ cells to hydroxyurea-induced replication stress. **a**, Co-IP of Dpb11 and Slx4 after 2 h treatment with 0.1 M hydroxyurea (HU). **b**, Model for the action of Slx4 and Rtt107 in *mrc1*Δ cells. **c**, HU treatment of *mrc1*Δ cells leads to H2A^{pS129} accumulation near an origin of replication. Chromatin immunoprecipitation (ChIP) analysis of H2A^{pS129} at neighbouring *ARS305* regions I, II and III in the indicated strains after exposure to 0.2 M HU. Data are presented as mean + s.e.m. ($n = 3$). *, significant difference ($P = 0.025$), as analysed by unpaired, two-tailed student's *t*-test. **d**, HU sensitivity assay on the indicated mutants. **e**, Western blot showing the phosphorylation status of Rad53–HA after treatment with 10 mM HU. kb, kilobases.

METHODS SUMMARY

The yeast strains and plasmids used are described in Supplementary Tables 3 and 4. The mass-spectrometry-based phosphorylation analyses of whole cell lysates or purified Rad53, as well as the other procedures used for cell growth and synchronization, genotoxin treatment and chromatin immunoprecipitation analysis, are detailed in the Methods.

Full Methods and any associated references are available in the online version of the paper.

Received 18 August 2011; accepted 4 October 2012.
Published online 18 November 2012.

1. Weinert, T. A. & Hartwell, L. H. The *RAD9* gene controls the cell cycle response to DNA damage in *Saccharomyces cerevisiae*. *Science* **241,** 317–322 (1988).
2. Clerici, M. *et al.* Hyperactivation of the yeast DNA damage checkpoint by *TEL1* and *DDC2* overexpression. *EMBO J.* **20,** 6485–6498 (2001).
3. Pellicioli, A., Lee, S. E., Lucca, C., Foiani, M. & Haber, J. E. Regulation of *Saccharomyces* Rad53 checkpoint kinase during adaptation from DNA damage-induced G2/M arrest. *Mol. Cell* **7,** 293–300 (2001).
4. Fekairi, S. *et al.* Human SLX4 is a Holliday junction resolvase subunit that binds multiple DNA repair/recombination endonucleases. *Cell* **138,** 78–89 (2009).
5. Svendsen, J. M. *et al.* Mammalian BTBD12/SLX4 assembles a Holliday junction resolvase and is required for DNA repair. *Cell* **138,** 63–77 (2009).
6. Muñoz, I. M. *et al.* Coordination of structure-specific nucleases by human SLX4/BTBD12 is required for DNA repair. *Mol. Cell* **35,** 116–127 (2009).
7. Fricke, W. M. & Brill, S. J. Slx1–Slx4 is a second structure-specific endonuclease functionally redundant with Sgs1–Top3. *Genes Dev.* **17,** 1768–1778 (2003).
8. Stoepker, C. *et al.* SLX4, a coordinator of structure-specific endonucleases, is mutated in a new Fanconi anemia subtype. *Nature Genet.* **43,** 138–141 (2011).
9. Kim, Y. *et al.* Mutations of the *SLX4* gene in Fanconi anemia. *Nature Genet.* **43,** 142–146 (2011).
10. Roberts, T. M. *et al.* Slx4 regulates DNA damage checkpoint-dependent phosphorylation of the BRCT domain protein Rtt107/Esc4. *Mol. Biol. Cell* **17,** 539–548 (2006).
11. Schwartz, M. F. *et al.* Rad9 phosphorylation sites couple Rad53 to the *Saccharomyces cerevisiae* DNA damage checkpoint. *Mol. Cell* **9,** 1055–1065 (2002).
12. Schwartz, M. F., Lee, S. J., Duong, J. K., Eminaga, S. & Stern, D. F. FHA domain-mediated DNA checkpoint regulation of Rad53. *Cell Cycle* **2,** 381–394 (2003).
13. Tercero, J. A., Longhese, M. P. & Diffley, J. F. A central role for DNA replication forks in checkpoint activation and response. *Mol. Cell* **11,** 1323–1336 (2003).
14. Ohouo, P. Y., Bastos de Oliveira, F. M., Almeida, B. S. & Smolka, M. B. DNA damage signaling recruits the Rtt107–Slx4 scaffolds via Dpb11 to mediate replication stress response. *Mol. Cell* **39,** 300–306 (2010).
15. Navadgi-Patil, V. M. & Burgers, P. M. Yeast DNA replication protein Dpb11 activates the Mec1/ATR checkpoint kinase. *J. Biol. Chem.* **283,** 35853–35859 (2008).
16. Mordes, D. A., Nam, E. A. & Cortez, D. Dpb11 activates the Mec1–Ddc2 complex. *Proc. Natl Acad. Sci. USA* **105,** 18730–18734 (2008).
17. Granata, M. *et al.* Dynamics of Rad9 chromatin binding and checkpoint function are mediated by its dimerization and are cell cycle-regulated by CDK1 activity. *PLoS Genet.* **6,** e1001047 (2010).
18. Pfander, B. & Diffley, J. F. Dpb11 coordinates Mec1 kinase activation with cell cycle-regulated Rad9 recruitment. *EMBO J.* **30,** 4897–4907 (2011).
19. Tanaka, S. *et al.* CDK-dependent phosphorylation of Sld2 and Sld3 initiates DNA replication in budding yeast. *Nature* **445,** 328–332 (2007).
20. Zegerman, P. & Diffley, J. F. Phosphorylation of Sld2 and Sld3 by cyclin-dependent kinases promotes DNA replication in budding yeast. *Nature* **445,** 281–285 (2007).
21. Puddu, F. *et al.* Phosphorylation of the budding yeast 9-1-1 complex is required for Dpb11 function in the full activation of the UV-induced DNA damage checkpoint. *Mol. Cell. Biol.* **28,** 4782–4793 (2008).
22. Li, X. *et al.* Structure of C-terminal tandem BRCT repeats of Rtt107 protein reveals critical role in interaction with phosphorylated histone H2A during DNA damage repair. *J. Biol. Chem.* **287,** 9137–9146 (2012).

23. Javaheri, A. *et al.* Yeast G1 DNA damage checkpoint regulation by H2A phosphorylation is independent of chromatin remodeling. *Proc. Natl Acad. Sci. USA* **103,** 13771–13776 (2006).
24. Alcasabas, A. A. *et al.* Mrc1 transduces signals of DNA replication stress to activate Rad53. *Nature Cell Biol.* **3,** 958–965 (2001).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work was supported by grants from the National Institutes of Health (RO1-GM097272 to M.B.S. and F31-GM093588 to P.Y.O.). F.M.B.O. was supported by a Cornell Fleming Research Fellowship. C.J.M. was supported by an HHMI Institutional Undergraduate Education Grant to Cornell. The authors thank B. Almeida for technical assistance and R. Weiss, S. Emr, A. Bretscher, G. Balmus and P. Russell for comments on the manuscript.

Author Contributions P.Y.O., F.M.B.O. and M.B.S. designed and performed experiments and analysed the data. P.Y.O. and M.B.S. performed the mass spectrometry experiments. F.M.B.O. performed the chromatin immunoprecipitation analysis and generated the *slx4* mutants. Y.L. and P.Y.O. performed co-immunoprecipitations between Dpb11 and Rad9. Y.L. performed pull-down experiments with the BRCT domains of Dpb11. C.J.M. performed the Rtt107–H2A binding assay and the experiments with the Rtt107 BRCT domains. P.Y.O. and M.B.S. performed experiments involving the overexpression of Slx4. P.Y.O. and M.B.S. wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.B.S. (mbs266@cornell.edu).

METHODS

Yeast strains and plasmids. Strains generated in this study were derived either from MBS164 or MBS191 (both congenic to S288C) or W303 (where indicated). Unless indicated, all tags were inserted at the C terminus of the corresponding genes by homologous recombination at the genomic loci and were verified by western blotting. Tagged strains were assayed for sensitivity to MMS to ensure they behaved similarly to the wild-type strain. Sensitivity assays were independently confirmed in strains derived from freshly sporulated diploids. Standard cloning methods were used to generate the plasmids for this study. Plasmids containing domains of Dpb11 or Rtt107 tagged at the amino terminus with a PATH tag (2× protein A + TEV cleavage site + 6× His)[25] were based on the pET21a vector (Novagen). HA-tagged full-length Rtt107, together with its native promoter, was cloned into pRS416 (Stratagene) to generate pMBS163. Wild-type alleles cloned into the pFA6a vector (Addgene) were linearized before integration into the respective endogenous loci. A pYES2/NT C vector (Invitrogen) containing full-length or C-terminal BRCTs of Rtt107 was used for overexpression. *SLX4* and *DPB11* constructs containing an *ADH1* or a *TDH3* promoter were generated by fusing the respective promoters (800 base pairs upstream of the start codon) to the corresponding open reading frame. The resultant PCR products were subsequently cloned into the pRS416 or pFA6a vector. All point mutations were generated by site-directed mutagenesis using either the QuikChange Multi Site-Directed Mutagenesis Kit (Stratagene) or the PFU Ultra II kit (Agilent). All yeast strains and plasmids used in this study are described in Supplementary Tables 3 and 4 and are available on request.

Cell synchronization and genotoxin treatment. Yeast cells were grown in yeast peptone dextrose (YPD) or drop-out medium at 30 °C. Log phase cultures (optical density at 600 nm ≈ 0.3) were subjected to α-factor (0.5 μg ml^{-1}) or nocodazole (1.5 μg ml^{-1}) treatment for G1 or G2/M arrest, respectively. Cells were then washed and resuspended in warm medium containing the indicated genotoxin.

Western blotting and immunoprecipitation. For western blotting, about 50 mg frozen cell pellet was lysed by bead beating at 4 °C in lysis buffer (50 mM Tris-HCl, pH 7.5, 0.2% Tergitol, 150 mM NaCl, 5 mM EDTA, 1 mM phenylmethylsulphonyl fluoride (PMSF), Complete, EDTA-free Protease Inhibitor Cocktail (Roche) and PhosSTOP (Roche)). SDS loading buffer with 60 mM dithiothreitol (DTT) was added. Samples were separated by standard SDS-polyacrylamide gel electrophoresis (SDS–PAGE). Proteins were detected using the following antibodies: anti-Rad53 (yc-19, 1:10,000, Santa Cruz Biotechnology), anti H2A^{pS129} (07-0745, 1:10,000, Millipore), anti-HA (12CA5, 1:10,000, Roche) and anti-Flag (M2, 1:5,000, Sigma) antibodies. For immunoprecipitation (IP), approximately 100 mg frozen cell pellet was lysed by bead beating at 4 °C in lysis buffer (50 mM Tris-HCl, pH 7.5, 0.2% Tergitol, 150 mM NaCl, 5 mM EDTA, 1 mM PMSF, Complete, EDTA-free Protease Inhibitor Cocktail, 5 mM sodium fluoride and 10 mM β-glycerophosphate). After adjusting protein concentrations to about 6 mg ml^{-1}, inputs were aliquoted, and lysates were incubated with either anti-HA or anti-Flag agarose resin (Sigma) for 2–3 h at 4 °C. After three washes in lysis buffer, bound proteins were eluted with three resin volumes of SDS elution buffer (100 mM Tris-HCl, pH 8.0, and 1% SDS) for HA IP or of Flag peptide (Sigma) solution (0.5 μg ml^{-1} in 100 mM Tris and 0.2% Tergitol) for Flag IP. SDS loading buffer with DTT was added, and samples were analysed by western blotting with the indicated antibodies.

Pull-down with recombinant BRCT domain. Protein domains (for Dpb11 BRCT$^{1/2}$, amino acids 1–270; and BRCT$^{3/4}$, amino acids 271–582) containing an N-terminal PATH tag (see Yeast strains and plasmids) were expressed in *Escherichia coli*, bound to human IgG–agarose resin (GE Healthcare) and then used as bait for pull-downs from yeast lysates as previously described[25].

SILAC labelling of yeast. For mass spectrometry experiments, cells were grown in (−)Arg (−)Lys drop-out medium ('light' version complemented with normal arginine and lysine; 'heavy' version complemented with lysine $^{13}C_6$, $^{15}N_2$ and arginine $^{13}C_6$, $^{15}N_4$) for at least five generations.

Purification of phosphopeptides by immobilized metal-ion affinity chromatography (IMAC). For the purification of Rad53 phosphopeptides, approximately 0.6 g cell pellet from wild-type (grown in heavy medium) or *slx4*Δ (grown in light medium) strains carrying Rad53–HA was lysed by bead beating at 4 °C in 4 ml lysis buffer (50 mM Tris-HCl, pH 7.5, 0.2% Tergitol, 150 mM NaCl, 5 mM EDTA, Complete, EDTA-free Protease Inhibitor Cocktail, 5 mM sodium fluoride and 10 mM β-glycerophosphate). Lysates were incubated with anti-HA–agarose resin (Sigma) for 4 h at 4 °C. After three washes with lysis buffer, bound proteins were eluted with three resin volumes of elution buffer (100 mM Tris-HCl, pH 8.0, and 1% SDS). Eluted proteins from light or heavy medium were mixed together, reduced, alkylated and precipitated. Proteins were resuspended in a solution of 2 M urea and 12.5 mM Tris-HCl, pH 8.0, and digested with trypsin for 16 h at 37 °C. Phosphopeptides were enriched using an 'in-house' IMAC column, then eluted with 10% ammonia and 10% acetonitrile and dried in a SpeedVac evaporator.

Phosphoproteome analysis. Approximately 0.6 g cell pellet from wild-type (grown in light medium) and *slx4*Δ (grown in heavy medium) strains was lysed by bead beating at 4 °C in 4 ml of lysis buffer (50 mM Tris-HCl, pH 7.5, 0.2% Tergitol, 150 mM NaCl, 5 mM EDTA, Complete, EDTA-free Protease Inhibitor Cocktail (Roche), 5 mM sodium fluoride and 10 mM β-glycerophosphate). Protein lysates were denatured in 1% SDS, reduced, alkylated and then precipitated with three volumes of a solution containing 50% acetone and 50% ethanol. Proteins were solubilized in a solution of 2 M urea, 50 mM Tris-HCl, pH 8.0, and 150 mM NaCl, and then TPCK-treated trypsin was added. Digestion was performed overnight at 37 °C, and then trifluoroacetic acid and formic acid were added to a final concentration of 0.2%. Peptides were desalted with a Sep-Pak C18 column (Waters), dried in a SpeedVac evaporator and resuspended in 1% acetic acid. Phosphopeptides were enriched by IMAC as previously described[26–28], reconstituted in 85 μl solution containing 80% acetonitrile and 1% formic acid, and fractionated by hydrophilic interaction liquid chromatography (HILIC) as previously described[26], before analysis by liquid chromatography with tandem mass spectrometry (LC-MS/MS). More than 3,570 phosphopeptides were identified or quantified, and phosphopeptides containing a phosphorylation site known to be a Mec1 target or a target of Rad53-dependent phosphorylation[27,29] were selected.

Mass spectrometry analysis. IMAC elutions or HILIC fractions were dried in a SpeedVac evaporator, reconstituted in 0.1% trifluoroacetic acid and analysed by LC-MS/MS using a 125 μM ID capillary C18 column and an Orbitrap XL mass spectrometer coupled with an Eksigent nanoflow system. Database searching was performed using the SORCERER system (Sage-N Research) running the program SEQUEST. After searching a target-decoy budding yeast database, results were filtered either based on probability score to achieve a 1% false positive rate or manual inspection. Quantification of heavy/light peptide isotope ratios was performed using the Xpress program as previously described[27].

Chromatin immunoprecipitation (ChIP). Cultures were grown in YPD to an optical density at 600 nm ≈ 0.3, arrested in G1 with α-factor for 2 h and released in the presence of 200 mM hydroxyurea (HU) for 1 h. Cultures were formaldehyde-crosslinked (1% final concentration) for 20 min followed by quenching with 125 mM glycine. Protein–DNA complexes were immunoprecipitated with a yeast-specific anti-phospho-histone (H2A^{pS129}) antibody (07-745, Millipore). Quantitative PCR was performed with the purified DNA from the immunoprecipitate as previously described[30] using pairs of primers designed to amplify the genome sequences shown in Fig. 4c. The primer sequences were as follows: 39.3Kb_for, 5′-CAAGTGG ATTGAGGCCACAGCA-3′, 39.3Kb_rev, 5′-CCGGACAGTACATGAAACT GGACA-3′; 43.1Kb_for, 5′-TCAAGGTGGCTTGATGATCGCC-3′, 43.1Kb_rev, 5′-CACCTCCAATCTGCTTCAAGTTTGGC-3′; 47.3Kb_for, 5′-TATCTTGCG GGCCTTTCGTGTC-3′, and 47.3Kb_rev, 5′-GGGAGATTCCATTTCCGCA CCA-3′.

25. Smolka, M. B. *et al.* An FHA domain-mediated protein interaction network of Rad53 reveals its role in polarized cell growth. *J. Cell Biol.* **175,** 743–753 (2006).
26. Albuquerque, C. P. *et al.* A multidimensional chromatography technology for in-depth phosphoproteome analysis. *Mol. Cell. Proteomics* **7,** 1389–1396 (2008).
27. Smolka, M. B., Albuquerque, C. P., Chen, S. H. & Zhou, H. Proteome-wide identification of *in vivo* targets of DNA damage checkpoint kinases. *Proc. Natl Acad. Sci. USA* **104,** 10364–10369 (2007).
28. Smolka, M. B. *et al.* Dynamic changes in protein–protein interaction and protein phosphorylation probed with amine-reactive isotope tag. *Mol. Cell. Proteomics* **4,** 1358–1369 (2005).
29. Chen, S. H., Albuquerque, C. P., Liang, J., Suhandynata, R. T. & Zhou, H. A proteome-wide analysis of kinase-substrate network in the DNA damage response. *J. Biol. Chem.* **285,** 12803–12812 (2010).
30. Petesch, S. J. & Lis, J. T. Rapid, transcription-independent loss of nucleosomes over a large chromatin domain at *Hsp70* loci. *Cell* **134,** 74–84 (2008).

CAREERS

TURNING POINT Biologist turned careers adviser offers pointers to scientists **p.126**

@NATUREJOBS Follow us on Twitter for the latest news and features **go.nature.com/e492gf**

NATUREJOBS For the latest career listings and advice **www.naturejobs.com**

MATT KENYON/GETTY

COLUMN

Leadership hurdles

Asian researchers and engineers are too rarely made US science leaders, say **Lilian Gomory Wu** and **Wei Jing**.

Many people believe that Asians excel in science, technology, engineering and maths (STEM) occupations in the United States. And indeed, there are lots of people of Asian descent on the country's university campuses and in its STEM workplaces and federal laboratories. In 2009, Asians — defined as people from the Far East, southeast Asia and the Indian subcontinent — made up 78% of doctoral recipients with temporary visas who were planning to work in the United States[1]. One could expect, then, that Asian men and women would be represented in similar proportions at the highest levels of STEM industry, academia and the federal workforce. But the data tell a different story.

Across all sectors, Asians in US STEM careers are not reaching leadership positions at the same rate as white people, or even as members of other underrepresented groups[2]. In academia, just 42% of Asian men are tenured, compared with 58% of white men, 49% of black men and 50% of Hispanic men. Just 21% of Asian women in academia are tenured, the lowest proportion for any ethnicity or gender. They are also least likely to be promoted to full professor.

The industrial and federal workforces reflect similar numbers. Asian men are doing better than Asian women in reaching managerial positions in industry, but their numbers are lower than those for men of other races and ethnicities. Just 4% of Asian women in industry and 28% in the federal workforce hold managerial positions, again the smallest percentage for any ethnicity or gender.

Asians are almost absent at the very top of US companies. The company Leadership Education for Asian Pacifics, based in Los Angeles, California, reported[3] in 2010 that there were just ten Asians or Pacific Islanders among the chairs, presidents and chief executives of the 500 biggest US firms; only three of them were women.

Why the disparity? It may be down to cultural behaviours, and Western interpretation of these behaviours. Asians are often stereotyped as a 'model minority': hardworking and patient, family oriented, good at maths and science and having a strong work ethic, but also humble, non-confrontational and lacking the passion to be charismatic leaders. Worse yet, a work group of the US government's Equal Employment Opportunity Commission reports[4] that Asians are often perceived as 'forever foreign', which can affect how others assess their ability to communicate, their competence and, more importantly, their trustworthiness.

Good leadership has a cultural dimension. In east Asia, for example, effective leadership is measured by what managers do rather than by what they say, no matter how passionately they speak. A manager in charge of bringing out a product there would work day and night to get it out on time and free of defects. Communication skills are generally less important in this model. The idea in the United States that east Asians lack passion and opinions comes from cultural perceptions of their behaviour: in discussions, east Asians tend to respond slowly, taking time to listen to what is being said ▶

and thus giving the appearance to Americans that they are not engaged, are passive and have no opinion. These differences can easily lead to unintended biases.

The problem may go beyond verbal communication. Grant applications to the US National Science Foundation from Asian principal investigators between 2004 and 2011 have been consistently funded in lower proportions than those from black, Hispanic and white principal investigators[5], which suggests that differences in writing styles may lead to biases. For example, east Asians' humble demeanour could cause them to describe the implications of their research in modest terms, which might bring them lower ratings from reviewers.

The idea of what makes a good leader in the United States needs to be re-examined. Cultural differences in communication style need further study; peer-review panels, managers and others should be trained to avoid biases. One model is the Strategies and Tactics for Recruiting to Improve Diversity and Excellence programme at the University of Michigan in Ann Arbor. Such programmes help scientists and engineers to be more effective in global collaborations and careers. At the same time, Asians need to recognize that hard work is not enough; they should seek training in communication, assertiveness and leadership skills.

The inequalities that mark the career arcs of Asian scientists and engineers in the United States are not widely discussed; the science community needs to bring greater attention to the data. We also need to look at whether Asians are recognized for their achievements, and whether they are receiving awards and becoming members of the US National Academies in numbers roughly equivalent to the proportion of Asians who rise to the level of full professor.

Diversity is said to be a strength of the United States. If cultural differences are recognized and respected, the country's scientific enterprise is sure to benefit. ■

Lilian Gomory Wu *is the programme executive of IBM University Programs Worldwide in Somers, New York.* **Wei Jing** *is a research associate in the Policy and Global Affairs division of the National Academies in Washington DC.*

1. *Doctorate Recipients from U.S. Universities: 2009* NSF 11-306 (National Science Foundation, 2010).
2. Wu, L. & Jing, W. *Issues Sci. Technol.* **28,** 82–87 (2011).
3. LEAP *2010 API Representation on Fortune 500 Boards* (LEAP, 2010).
4. *Asian American and Pacific Islander Work Group Report to the Chair of the Equal Employment Opportunity Commission* (EEOC, 2008).
5. *Report to the National Science Board on the National Science Foundation's Merit Review Process Fiscal Year 2011* (National Science Foundation, 2012).

TURNING POINT
Sarah Blackford

Science-careers adviser Sarah Blackford, head of education and public affairs at the Society for Experimental Biology in Lancaster, UK, assumed that she would be a research scientist. But after she landed a contract-research post, she realized that her interests lay elsewhere, and she manoeuvred through a series of jobs from journal publishing to careers development. In October, Blackford published her first book, Career Planning for Research Bioscientists *(Wiley-Blackwell). She is on the steering committee for the* Naturejobs *Career Expo.*

What did you hate about research?
I used to find it really tough doing the experiments. I am just not a very practical, technical person, and don't follow protocol very well — I can't go by a recipe in the kitchen.

But were there aspects that you enjoyed?
Presenting results in papers and posters, and going to conferences. I also liked interacting with people — negotiating for equipment, for example. When my contract came to an end, I thought about scientific publishing. I would have a foot in science, but would not be doing lab work. Everyone breathed a sigh of relief.

How did you transfer to careers advising?
I was the assistant editor at the *Journal of Experimental Botany*, based at the University of Southampton, UK, and biologists there kept bringing me their CVs — because I worked in publishing, they thought I would be a guru on language and writing. I enjoyed helping them, so I started volunteering at the university's career-services centre. I helped with CV workshops and sat in on interviews.

How did you move into paid careers advising?
My job relocated to Lancaster University when the journal editor changed, so I went to volunteer with their career services — and this was the turning point for my life. They were looking for someone to cover for a person on sabbatical, and I got the job. It was for only three months, but I knew I had to take it — and as it turned out, the job lasted for two years. After that, I had enough experience to get a job at the University of Leeds, UK, for a year and a half, where I wrote marketing plans, organized conferences, liaised with employers, ran careers workshops.

What prompted you to write a book?
I missed working with scientists, and a job came up with the Society for Experimental Biology involving career development, science communication and education. I have been in the post since 1998. A few years ago, while running careers workshops at a conference in Finland, I was chatting with a marketing manager and said flippantly that one day I would put all this information into a book. When I got back to my office, I had an e-mail from the commissioning editor at a publishing house saying he understood that I was thinking about writing a book.

Why did you focus on bioscience careers?
All this valuable careers information is being directed to people at conferences, but there was almost nothing in writing for bioscientists.

Do you have advice for biomedicine postdocs?
They need to keep learning new techniques and skills. They need to campaign for better contracts, the right to develop management skills, the opportunity to teach or do whatever they want to do to improve their career prospects. They can't let their supervisor steer for them. Universities are employing fewer technicians now, and postdocs are in danger of becoming supertechs. They also need to decide whether taking a third postdoc is an advantage. It may be convenient, but they ought to ensure that it will build on their current capabilities so that they are improving their career prospects.

What caveats do you find yourself repeating to early-career biomedical researchers?
You have to sell yourself. One of the easiest ways is through social media and networks. You need to network, because it is other people who get you jobs. Postdocs especially aren't using social media and networks enough: LinkedIn, for example, is extremely valuable because a lot of recruiters use it. You can meet influential people online — modern networks are very democratic. Opportunities are out there. ■

INTERVIEW BY KAREN KAPLAN

FIRST FOOT

Everything as it should be.

BY DEBORAH WALKER

JACEY

New Year's Eve and it's snowing outside. Of course it is. They switched on the weather machines on 1 December. Snow for the holidays. We're in the living room, waiting for the first foot. The first person through the door who will bring our luck for the year.

Uncle Milo's dozing in his chair. He's as drunk as a skunk. He was never a heavy drinker. But we all have our quotas.

Uncle Milo and Auntie Val live with us. They used to have a nice retirement flat overlooking the harbour. But families should live together.

The lights are dim, the red fairy lights on the Christmas tree are winking and recording, checking that the traditions are being upheld. I wonder if any observers are tuned in. Are we providing good entertainment?

I go to the window, stare out into the street. There's a hundred families, a hundred houses in a circle. A hundred houses with no back doors or windows. There are lights in every sitting room. We're all waiting for our first foot. I can see them, tall dark figures in the cold, waiting to bring our luck.

With a peal of bells from a non-existent church, the year turns.

"Happy New Year," shouts Mother. When the twins jump up and down, shouting in excitement, Mother looks relieved. She'd spent all afternoon coaching them. I smile at her, trying to tell her, that I understand. I really do.

She pats my cheek. "Happy New Year, Brenna."

Auntie Val rouses Uncle Milo. We link hands and sing *Auld Lang Syne*.

Mother looks towards the door. "Where is he? Where is he?" she whispers.

"It's all nonsense." Uncle Milo's eyes are red and bleary. He strides towards the Christmas tree and glares at it. "It's all nonsense." His words fall like the heavy snow.

I pull the twins closer to me. They're still bewildered, after seeing Santa delivering the presents. Santa shouldn't ooze down the chimney. Sometimes the enforcers get things wrong. But it's no good complaining.

"Is it going to be all right?" asks Corey. Jane buries her face into my shoulder. *No*, I feel like saying. *It's not going to be all right. How can it be?* But it's not fair to them. I smile and tell them that it's going to be fine. They should be allowed to hope. They're only kiddies.

Auntie Val is trying to calm Uncle Milo. "Hush now," she says, laying a hand on his shoulder. Auntie Val looks so frail. Since the abduction, she's grown smaller and smaller, despite all the food she's had to eat during the season's celebrations. I'm worried about Auntie. I'm worried about us all. I'm all worn out with the worry. Worry is a knife, and it's whittled me hollow.

"I won't be hushed." Uncle Milo's voice grows louder, accumulating like a rolling snowball. "We. Can't. Live. Like. This. You can't force us to enjoy ourselves." He takes a swing at the tree. The red fairy lights flash. Uncle stumbles. Mother gives a little gasp. We both reach out to try to catch him. But we're too slow. He falls to the floor, landing badly on the wooden abattoir that Santa bought Corey for Christmas.

"I'm okay. I'm okay," he says. "Don't fuss so. Don't," he says, shaking off Auntie Val's offer of help. He sounds as if he hates her.

Auntie Val begins to cry. So do the twins.

"Unseasonal behaviour won't be tolerated." The enforcer's voice fills the room.

Mother says quickly, "But it's traditional for people to lose their rag at this time of year."

"Ensure that it's an isolated incident."

"I will, thank you."

There's a knock at the door.

"Answer it, Brenna," says Mother.

I open the door. He's tall and he's dark. Father smiles at me, but he looks puzzled. I feel uneasy. Usually Father's so good at playing his part.

"Where were you?" hisses Mother. "Uncle Milo's had a turn."

Uncle's still on the floor, sobbing.

"I… met someone."

"Just get it over with," says Mother.

Father holds a handful of silver coins, a lump of coal and a twist of salt. He shouts out: "Happy New Year t'ye! God send ye plenty! Where ye have one pound note, I wish ye have twenty." He passes out the gifts, giving me the salt twist.

I unwrap it and taste the salt with the tip of my finger. Then I let the salt fall to the floor, and quickly shove the wrapper into my pocket. The paper's printed with a red circle: the sign of the resistance. I'd heard rumours but I never thought they could be true. Can anyone fight the enforcers? Can anyone escape this zoo?

Mother gives Father a mince pie. It's dusted with blue sugar. Father looks at it for a moment, before eating it with two quick bites.

Mother sighs. "That's it, then. Let's get to bed."

"Happy New Year, Mother." I hug her.

"Happy New Year, darling."

I help the twins to bed, thinking about the first foot who brings the luck for the New Year. And I'm thinking about the resistance. The New Year brings hope. ■

Deborah Walker *grew up in the most English town in the country, but she soon high-tailed it down to London, where she now lives with her partner, Chris, and her two young children.*